THESE

présentée à

L'UNIVERSITE DE PARIS VII

pour l'obtention du

DOCTORAT D'ETAT ES SCIENCES

par

Anne-Marie DEROUAULT

Modélisation d'une langue naturelle pour la désambiguation des chaînes phonétiques

soutenue le 26 avril 1985 devant le jury composé de:

A. LENTIN
M. GROSS
D. PERRIN
M. NIVAT
R. MOREAU
F. JELINEK

MODELISATION D'UNE LANGUE NATURELLE POUR LA DESAMBIGUATION DES CHAINES PHONETIQUES.

Anne-Marie Derouault

Maurice GROSS, directeur du Laboratoire d'Automatique Documentaire et Linguistique, a bien voulu accepter de diriger cette thèse. Je tiens à le remercier pour sa disponibilité, et pour les remarques et les suggestions toujours pertinentes qu'il m'a faites au long de ces trois années.

Je remercie vivement André LENTIN et Dominique PERRIN, d'avoir examiné cette thèse et accepté d'en être rapporteurs, ainsi que Maurice NIVAT qui a bien voulu participer au jury.

Ce travail a pu prendre forme grâce à l'intérêt que René MOREAU, directeur du Développement Scientifique d'IBM France, a toujours manifesté pour les sujets touchant les mathématiques et le langage. Dans cet effort pour amener le langage naturel à se laisser prédire, sinon mettre en équation, sa confiance et son soutien ne se sont jamais démentis. C'est grâce à son enthousiasme que tous les moyens nécessaires ont été mis à notre disposition. Qu'il me soit permis ici de lui exprimer ma profonde estime et ma gratitude.

Frederick JELINEK, qui dirige l'équipe de reconnaissance de la parole au laboratoire de recherche T. J. Watson d'IBM USA à Yorktown, m'a accueillie quelque temps dans son équipe. C'est à lui que je dois la découverte des applications de la théorie de l'Information au décodage linguistique. Je veux ici le remercier encore pour l'enseignement qu'il m'a donné, le temps passé ensemble à parler... de parole, et l'honneur qu'il me fait en acceptant de participer au jury de cette thèse.

Je souhaite également souligner que, dans l'équipe formée avec Bernard MERIALDO, l'interaction a été permanente, la collaboration enrichissante et stimulante. C'est, entre autres, grâce à l'entente établie, que ce travail a été, aussi, un plaisir.

Enfin, j'associe à ces remerciements les membres du Centre Scientifique, avec qui j'ai pu avoir de nombreux échanges de points de vue, et particulièrement son directeur, Alain CROISIER, pour les excellentes conditions de travail et de rédaction dont j'ai profité.

Les membres de l'équipe de Yorktown m'ont toujours accueillie avec gentillesse et une grande disponibilité. Je leur exprime mes remerciements cordiaux.

TABLE DES MATIERES

PREFACE

Les progrès réalisés dans le matériel informatique font de l'ordinateur un outil susceptible d'être utilisé dans un nombre croissant de domaines. Pour les utilisateurs finaux, la communication homme-machine est un facteur clé du succès des applications. Celle-ci doit être le plus commode possible, souple et agréable. De plus, la sophistication attendue des machines de la cinquième génération demande que cette communication possède en apparence les traits de la communication humaine. Si cela se réalise même partiellement, elle peut ouvrir un champ d'applications encore inexploré. Compréhension du langage naturel, intelligence artificielle, entrée et synthèse vocales sont ainsi actuellement des thèmes de recherche extrêmement féconds. Cependant nous savons que nous sommes encore très loin du but ultime que serait la communication dans les deux sens en langage naturel, par la parole continue. Entre autres, la reconnaissance de parole soulève de vastes problèmes à la fois d'ordre acoustique et d'ordre linguistique:

Partant du signal de parole, un système de reconnaissance doit d'abord extraire des unités telles que les phonèmes, ou les syllabes. Cette phase relève du traitement de signal et donne actuellement des taux de performance autour de 60-70%.

Cette performance médiocre explique que les systèmes de reconnaissance existants limitent la taille du vocabulaire, ou restreignent la syntaxe des phrases permises. La phase suivante de décodage linguistique est alors facilitée et compense même

certaines erreurs du traitement acoustique, pour arriver en sortie à un taux de reconnaissance bien meilleur sur les mots. Cette phase consiste à retrouver les mots prononcés à partir des unités acoustiques reconnues.

De tels systèmes conviennent pour des applications dans un domaine très limité (commande vocale, réservations, tri par exemple). La taille du vocabulaire est alors typiquement de quelques centaines de mots.

Cependant il serait souhaitable pour des tâches de plus grande envergure de pouvoir traiter le discours naturel sans contrainte. Quelques recherches s'orientent donc vers la reconnaissance de phrases construites librement sur des vocabulaires plus larges (plus de mille mots). Le plus gros système actuel traite 5.000 mots anglais ([8])
Les problèmes linguistiques se posent alors à trois niveaux:
- lexical: choix des mots et des informations à stocker pour chacun,
- syntaxique: les phrases reconnues devraient être "correctes" d'un point de vue grammatical,
- sémantique enfin: comment repérer la cohérence de sens dans la phrase, et éviter de produire une phrase telle que "le moi de mais"...
C'est volontairement que le niveau pragmatique n'est même pas cité. En effet nous n'avons encore aucun embryon d'outil qui permettrait de distinguer "Je vous expose ceci avec le plus grand des plaisirs / déplaisir"...

Notre travail a porté sur ces problèmes de décodage linguistique à partir d'une entrée phonétique de bonne qualité. Nous avons ainsi isolé le décodage linguistique pur du traitement acoustico-phonétique. Cela permet d'élaborer des modèles aussi généraux que possible, qui tiennent compte de difficultés habituellement occultées

lorsque l'on veut un système de reconnaissance opérationnel, à savoir:

- le vocabulaire est illimité.
- le domaine de discours est ouvert.

Comme bénéfice secondaire de ce travail, la possession d'un dictionnaire complet donne éventuellement la possibilité d'en étudier la complexité, de le comparer à d'autres langues, de le réduire pour passer à des applications pratiques plus limitées.

Un modèle général dont on connaît les performances et les limites peut ensuite être adapté, réduit ou autre pour une tâche particulière de reconnaissance. On obtient ainsi des modèles facilement extensibles par nature dès que l'on progresse suffisamment du côté acoustique pour élargir le domaine d'application ou le vocabulaire.

L'entrée phonétique de bonne qualité choisie ici est la transcription sténotypée du discours.

Plutôt que d'effectuer une simulation complète avec un programme de phonétisation, cette démarche donne une application utilisable par les professionnels de la sténotypie. Grâce à eux nous avons pu tester le système en situation réelle, et juger de l'acceptabilité du système en fonction des performances.

De plus, nous verrons que l'entrée sténotypée est en réalité pseudo phonétique, puisque certains phonèmes sont systématiquement confondus. Si elle est tout de même loin d'une phonétique avec erreurs produite par un processeur acoustique, elle est cependant plus floue et ambiguë qu'une phonétique obtenue par programme.

Nous décrirons dans une première partie les principes de la sténotypie, et le genre d'ambiguïtés qui en résulte pour la transcription. Nous ferons l'étude du dictionnaire sténotypie-français sur lequel s'appuie la transcription. Nous rappellerons les différentes stratégies qui s'offraient pour lever les ambiguïtés.

La deuxième partie exposera le modèle markovien du français et le système de transcription qui s'appuie sur ce modèle tel qu'il a été développé dans l'équipe du Centre Scientifique et implémenté sur un micro ordinateur. Enfin, dans la troisième partie, nous comparerons cette approche avec une méthode purement grammaticale. Nous montrerons comment une combinaison convenable des deux modèles permet d'obtenir de meilleurs résultats que chaque modèle seul.

PARTIE I: DEFINITION ET COMPLEXITE DE LA TACHE.

1. POSITION DU PROBLEME.

1.1. DEFINITION DE L'ENTREE PSEUDO-PHONETIQUE.

L'entrée phonétique que nous avons choisie est la transcription sténotypée du discours oral. Aussi commencerons-nous par rappeler les principes de la sténotypie.

La sténotypie a pour but de saisir la parole à la vitesse d'élocution. Elle est principalement utilisée pour la saisie en temps réel de textes de conférences, dans les domaines les plus divers (médecine, politique, techniques...). Comme la sténographie, elle fonctionne suivant un principe phonétique, c'est-à-dire que seule la prononciation entendue est prise en compte, mais absolument pas l'orthographe. Cette saisie s'effectue sur un clavier spécial.
En France, la seule méthode utilisée est la méthode Grandjean [1]. Le clavier comprend 21 touches, qui provoquent chacune l'impression d'un caractère sur une bande de papier. Cette bande comporte 21 colonnes, et un caractère donné s'imprime toujours dans la même colonne. Comme il y a en français environ 36 phonèmes, chaque touche sert à coder plusieurs phonèmes.

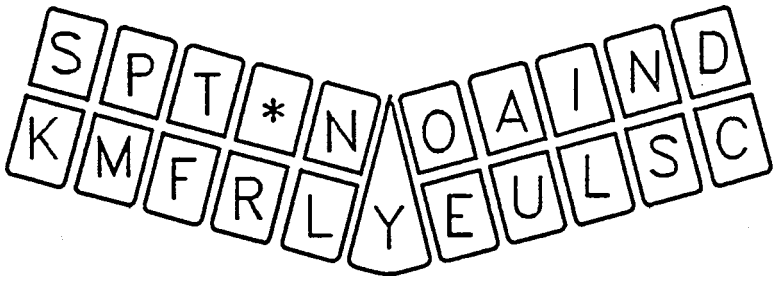


Figure 1: Le clavier de sténotypie Grandjean

Une frappe consiste à enfoncer simultanément plusieurs touches, ce qui provoque l'impression des caractères dans les colonnes correspondantes, sur une même ligne. Le fait de relâcher les touches provoque l'avancement de la bande de papier à la ligne suivante. Les touches sont réparties sur le clavier de façon à ce qu'une frappe puisse correspondre facilement à une syllabe:

- Les 11 touches de gauche (de S à Y) correspondent aux consonnes initiales, avec confusions des sons voisés et non voisés:

  "S"=se, ze "P"=pe, be "T"=te, de "K"=que, gue "F"=fe, ve.

  Les groupes consonnantiques br, bl, tr... sont obtenus en appuyant simultanément sur plusieurs touches ("PR", "PL", "TR"...).

  La touche "Y" sert à la fois pour le son je et pour la semi-voyelle y.

- Les 5 touches du milieu O ,A, I, E, U servent à coder les voyelles.

  "E"=é, è

  "A"=a, â.

"O"=o, ô.

Le e muet est omis.

Des combinaisons sont permises: "EU" pour eu, "OU" pour ou, "OI" pour oi, "AI" pour aille, etc.

Les nasales sont obtenues en ajoutant le "N" final.

- Les 5 touches de droite (de "L" à "C")[1] codent les consonnes finales avec les confusions:

  "L"=le, re "N"=nasales "S"=se, ze, fe, ve "D"=te, de "C"=que, gue.

De plus, des conventions autorisent certaines contractions de syllabes, (exemple: E TAD = étable) et définissent l'écriture de syllabes usuelles, dans le but de diminuer le nombre de frappes nécessaires, et d'augmenter la vitesse de saisie. De même, on peut contracter un mot d'une syllabe avec un "e" muet, comme "le", "de", "que" etc. avec la première syllabe du mot suivant chaque fois que cela est compatible avec l'ordre des touches du clavier. Par exemple: TLA="de la", KL="que le", etc.

Les coupures syllabiques sont laissées à l'appréciation du sténotypiste, ce qui produit éventuellement plusieurs sténos possibles pour un même mot:
Par exemple, le mot "rester" peut être coupé:

res/té et sténotypé RES TE,

---

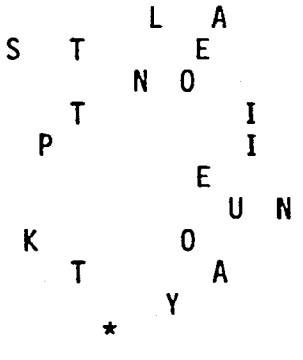
1 Sur la bande de papier, le N final est en réalité imprimé comme un N renversé, le S final comme un S barré, le C final comme un crochet.

ou bien ré/sté et sténotypé RE STE.

Bien sûr, des mots longs plurisyllabiques pourront avoir jusqu'à 8 sténos différentes.

Il n'y a pas de coupure entre les mots, ni en général de virgules. Seules les fins de phrases sont marquées, par la frappe "*" représentant le point.

La bande sténo que nous prenons comme entrée pour une transcription en français se présente donc comme une suite continue de syllabes pseudo-phonétiques:



la
ste-
no-
ti-
pie
est
un
co-
da-
ge
.

1.2. SOURCES D'AMBIGUITES POUR LA TRANSCRIPTION AUTOMATIQUE.

S'il est facile pour un humain de relire une bande sténo avec un peu d'habitude, le passage automatique de la bande aux mots correspondants pose des problèmes analogues à la transcription d'une suite de phonèmes en texte écrit, comme nous l'avons signalé en Introduction. La sténo est en effet très ambiguë, spécialement en français. Ceci est dû à trois facteurs:

- Les confusions systématiques entre phonèmes:

  Par exemple les mots pôle, Paul, Paule, porc, pore, bord, bol seront sténotypés de la même façon.

- Le grand nombre d'homonymes en français:

  soit des homonymes accidentels comme fond, font ou seau, sot, sceau

  soit des homonymies dues aux inflexions grammaticales d'un même mot comme donné, donnés, donnée, données, donnait, donnais, etc.

- L'absence de marque de fin de mot:

  un mot plurisyllabique peut en général être coupé en plusieurs mots plus courts prononcés de la même façon: par exemple "efficacité" et "et fit cas si thé"...

Etant donnée une bande sténo, on trouve en moyenne pour chaque syllabe 10 mots dans le dictionnaire pouvant couvrir cette syllabe et une ou plusieurs des syllabes suivantes (voir exemple page suivante). Le nombre de transcriptions possibles pour une seule phrase sténotypée augmente donc exponentiellement avec le nombre de mots.

La page suivante montre tous ces mots trouvés dans le dictionnaire pour la phrase "Qu'en est-il des obstacles qui s'opposaient à notre marche quand nous avons pris la route". La phrase du haut est la transcription choisie par le système que nous décrivons dans la suite de cet exposé.

Quand est il des obstacles qui s'opposaient à notre marche quand nous avons pris la route

KAN	Caen, quant, quand gants, gant, camps, camp, qu'en, qu'ans, qu'an.
E	hait, hais, haies, haie, et, est, es, ait, ais, aient, aie, ai.
IL	ils, il, îles, île.
TE	thés, thé, tait, tais, taies, taie, tes, des, dès, dés, dé, d'hait, d'hais, d'haies, d'haie, d'et, d'est, d'es, d'ait, d'ais, d'aient, d'aie, d'ai, t'hait, t'hais, t'haies, t'haie, t'et, t'est, t'es, t'ait, t'ais, t'aient, t'aie, t'ai.
OD	obstacles, obstacle, Aude, odes, ode, hautes, haute, hôtes, hôte, aubes, aube, ôtes, ôtent, ôte.
STAC	ce dagues, ce dague, se dagues, se dague.
KI	qui, gui, qu'y.
SO	sots, sot, saut, sauts, sceau, sceaux, seau, seaux, c'os, c'hauts, c'haut, c'eaux, c'eau, c'aux, c'au, s'os, s'hauts, s'haut, s'eaux, s'eau, s'aux, s'au, c'opposez, c'opposer, c'opposait, c'opposais, c'opposaient, c'opposés, c'opposées, c'opposée, c'opposé, s'opposez, s'opposer, s'opposait, s'opposais, s'opposaient, s'opposés, s'opposées, s'opposée, s'opposé.
PO	Pau, pots, pot, peaux, peau, beaux, beau, posez, poser, posait, posais, posaient, posai, posés, posées, posée, posé, bossé.
SE	ses, sait, sais, saie, c'hait, c'hais, c'haies, c'haie, c'et, c'est, c'es, c'ait, c'ais, c'aient, c'aie, c'ai, s'hait, s'hais, s'haies, s'haie, s'et, s'est, s'es, s'ait, s'ais, s'aient, s'aie, s'ai.
A	as, a, à, anneaux, anneau.
NO	nos, n'os, n'hauts, n'haut, n'eaux, n'eau, n'aux, n'au, notre, nôtres, nôtre, n'autres, n'autre.
TR	de remarche, te remarche.
MALS C	marches, marchent, marche, m'arches, m'arche.
KAN	Caen, quant, quand, gants, gant, camps, camp, qu'en, qu'ans, qu'an.
NOU	nous, noues, nouent, noue, n'houa, n'ou, n'où, n'houx, nouas, noua.
A	as, a, à, avons
FON	vont, font, fonds, fond.
PRI	prix, prit, pris, pries, prient, prie.
LA	las, la, là, l'as, l'a, l'à.
ROUD	routes, route.

2. LE DICTIONNAIRE.

2.1. CONTENU.

L'origine du dictionnaire est une liste de 150.000 formes fléchies, dont nous disposions et qui provenait de travaux antérieurs du Centre Scientifique. Chaque forme était rattachée à sa racine de la façon suivante:

substantifs avec la forme singulier,
adjectifs avec la forme masculin singulier,
verbes conjugués avec la forme infinitif.

De plus, chaque forme était accompagnée de sa fréquence dans un ensemble de textes français (utilisés pour extraire justement cette liste), et rattachée à l'une des six parties du discours: nom, adjectif, adverbe, préposition, pronom, verbe.

Pour réaliser un système de transcription, nous avons rajouté des informations indispensables: la sténotypie de chaque mot et des informations grammaticales plus détaillées.

2.1.1. Sténotypisation.

Chaque mot a été d'abord sténotypé par un programme (B.Mérialdo) utilisant:

- une décomposition automatique en syllabes [35]
- un programme de phonétisation [40]
- les règles de sténotypie [1]

Tandis que le modèle de langage était développé et testé en simulation sur des textes sténotypés automatiquement, le dictionnaire sténotypé était vérifié manuellement par des sténotypistes professionnels (un an de travail).[2]

2.1.2. Informations grammaticales.

Pour un si grand nombre de mots, il est nécessaire d'opérer une classification grâce à des informations grammaticales plus détaillées que les six parties du discours. Ces informations devaient au moins permettre de distinguer entre les inflexions homonymes d'un même mot afin de choisir la transcription juste dans un contexte donné. C'est pourquoi 92 classes morphologiques ont été retenues, permettant au maximum de désambiguïser la sténo mais pouvant aussi raisonnablement être attribuées aux 150.000 mots quasi automatiquement. Ce système de classes est donné in extenso dans "Annexe A. Liste des classes grammaticales" page 115.

---

2 Les variantes sténo d'un même mot ont été également rajoutées.

Essentiellement il comprend:

- quatre classes de substantifs et quatre d'adjectifs correspondant aux quatre couples (genre, nombre).
- 8 classes de pronoms personnels avec leur personne et genre, et une de pronoms objet.
- pour les verbes, autant de classes que de personnes (6 classes), plus une classe pour les infinitifs, une pour les participes passés et une autre pour les participes présents. Les auxiliaires sont aussi séparés, être ou avoir, avec la personne s'ils sont conjugués.

Ces classes détaillées ont été attribuées automatiquement à partir des informations d'origine, chaque fois que c'était possible.

2.1.3. Complétion.

Les 150.000 formes de ce dictionnaire ont été récemment vérifiées et complétées par un groupe du Centre Scientifique IBM travaillant sur la constitution d'un thésaurus ([22]). Les mots nouveaux ont été sténotypés par une professionnelle, et notre dictionnaire sténotypie-français comprend maintenant 250.000 formes fléchies.

Les expériences relatées dans la suite de cet exposé ont été effectuées avec la liste de 150.000 mots, étant donnée la date récente où la liste complète a été achevée. Cependant nous avons testé le système de transcription sur IBM PC avec le dictionnaire total, et les résultats annoncés dans la partie II restent valables, aussi

bien en ce qui concerne le taux d'erreurs que la vitesse de traitement. Ceci n'est pas surprenant si l'on remarque que, d'une part, les mots rajoutés sont généralement peu fréquents, longs et peu ambigus, et que d'autre part, l'organisation du dictionnaire, optimisée comme il est expliqué au paragraphe suivant, permet qu'un nombre important de mots soient rajoutés sans perdre beaucoup de temps de calcul.

2.2. ORGANISATION.

La recherche dans le dictionnaire doit permettre de trouver tous les mots dont la sténotypie figure sur la bande à transcrire. Le dictionnaire a donc été mis sous forme d'enregistrements contenant une sténo et les formes fléchies correspondantes avec leurs classes grammaticales et leurs fréquences.

La recherche élémentaire consiste à trouver tous les mots dont la sténo coïncide avec le début de la bande à transcrire: par exemple, pour la sténo "NOU A FON", trouver: nous, noue, noues, nouent, noua, nouas, nouât...
Nous appelons ces mots des "coïncidences courtes" (en anglais "short match") de la suite totale de syllabes. Une "coïncidence maximale" ("longest match") est un mot dont la sténo est la plus longue.

L'organisation du dictionnaire sténo-français a été optimisée en s'inspirant de la méthode suivie pour trouver en anglais la coïncidence maximale [36]. En classant les sténos par ordre lexicographique décroissant, et en rajoutant dans le même enregistrement toutes les coïncidences courtes de la sténo de cet enregistrement, il est possible de trouver en un seul accès disque, en même temps, toutes les coïncidences

courtes d'une bande donnée. Un tel dictionnaire a été implémenté par B. Mérialdo [16].

2.3. COMPLEXITE.

Après avoir passé en revue les différents types d'ambiguïtés dans la sténo, il est intéressant d'évaluer quantitativement la complexité inhérente au dictionnaire français tel que nous l'avons maintenant constitué, du point de vue de la transcription ou de la reconnaissance de parole. Nous pouvons aussi comparer cette complexité avec celle de l'anglais.

La description de la sténotypie donnée en I.1.1 montre qu'elle se déduit de la forme phonétique moyennant des marques de syllabes et une classification des phonèmes en 24 catégories:

10 catégories de consonnes initiales (confusion voisée/non voisée),

5 catégories de voyelles (confusion ouverte/fermée) et 4 nasales,

5 catégories de consonnes finales (nasale, sonorante, fricative, occlusive sauf palatale, palatale).

2.3.1. Mesures.

Dans ce paragraphe, nous étendrons le sens du mot "homonyme" aux mots ayant même sténo.

- Le nombre de sténos différentes est de 68.000. Il y a donc en moyenne 2,1 forme par sténo, avec un maximum de 37.

  Presque la moitié des formes (57.000) ont zéro, un ou deux homonymes. Ceci pourrait laisser croire que la tâche de la transcription ou de la reconnaissance est facile. En réalité, cela suppose que les frontières de mots sont connues, ce qui n'est le cas ni en sténotypie ni en parole continue. De plus, la plupart de ces mots peu ambigus sont longs et peu fréquents: si l'on calcule le nombre moyen de formes par sténo en pondérant cette fois par la fréquence des formes, on trouve 5 formes par sténo. D'ailleurs si l'on estime qu'il y a en moyenne 2 syllabes par mot, cela s'accorde avec le résultat expérimental de 10 mots pouvant commencer à chaque syllabe sténo (I.1.2).

- Les homonymes vraiment difficiles à distinguer en s'appuyant sur des informations uniquement grammaticales sont ceux qui ont une classe grammaticale commune. Ceux-ci réclameraient des informations sémantiques. Ils représentent 9% du dictionnaire.

- Il est d'autant plus difficile de déterminer les frontières des mots que le nombre de mots monosyllabiques est grand: notre dictionnaire en contient 6.000 (une syllabe sténo). A titre de comparaison, le dictionnaire du LADL (Laboratoire d'Automatique Documentaire et Linguistique) contient 3.000 mots d'une syllabe. La différence entre ces deux nombres s'explique cependant. En effet,

  le dictionnaire "DLAS" ([25]) du LADL ne contient pas les formes déclinées, mais les racines.

  le facteur de multiplication n'est cependant que de deux, car si dans les deux cas,

les mots comprenant une syllabe plus une consonne suivie d'un "e" muet, comme "pare", "toile" etc. comptent pour une seule syllabe, par contre les mots comme "tendre", "mordre" etc. avec deux consonnes et un e muet, sont comptés pour une syllabe pour le LADL, et pour deux syllabes sténos dans notre dictionnaire. Les 6.000 mots correspondent à 1.600 sténos différentes (facteur 4 au lieu de 2), et sont très ambigus. Une forte proportion (1.300) d'entre eux se trouve dans les 5.000 mots les plus fréquents. Ceci est bien sûr un trait spécifiquement français. Il n'est pas étonnant que 80% des sténos du dictionnaire puissent être coupées en plusieurs sténos plus courtes.

2.3.2. Comparaison avec l'anglais.

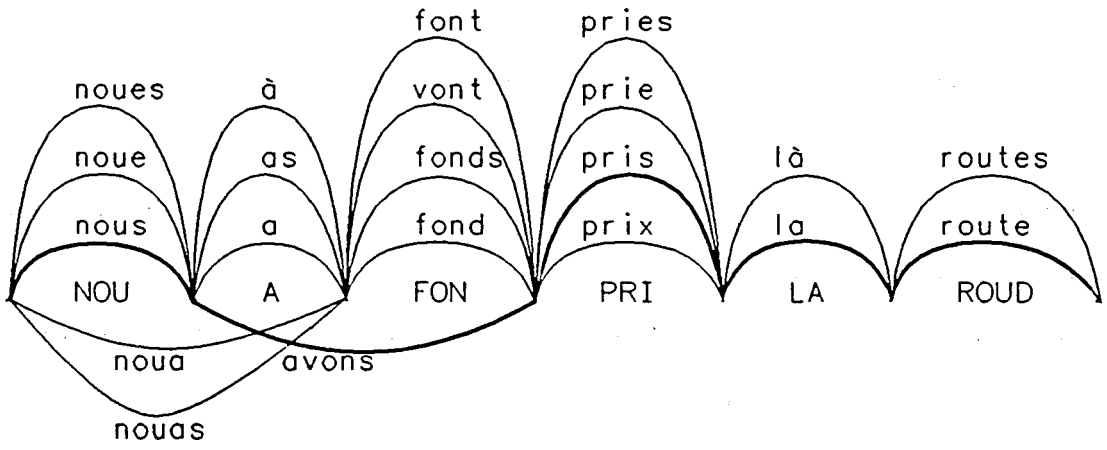
Une étude de V.Zue [43] sur un dictionnaire de 20.000 mots anglais classés d'après une représentation phonétique encore plus grossière que la nôtre (6 classes de phonèmes), sans marque de syllabes, indique que le nombre moyen de mots par représentation est aussi 2. Or nous utilisons une classification des phonèmes avec 4 fois plus de classes et trouvons le même nombre. Ceci montre que le français est significativement différent. En fait il paraît plus complexe (à cause de ces ambiguïtés) pour la reconnaissance de parole, dans le passage phonèmes-texte. Pour la synthèse de parole à partir du texte au contraire, le passage texte-phonèmes se décrit assez facilement avec des règles et quelques exceptions en français, mais pose des problèmes en anglais où certaines graphies se prononcent de beaucoup de manières différentes: "ough" par exemple est prononcé différemment dans "tough", "dough", "rough", "cough", etc.

D'autre part, les systèmes de transcription de la sténotypie anglaise basés simplement sur le principe du "longest match" obtiennent des performances satisfaisantes. Le mot le plus long est en général le bon. Les problèmes de frontières de mots et d'homonymie sont donc pratiquement inexistants: les inflexions sont beaucoup moins nombreuses (3 pour les verbes contre 40 en français) et non homonymes. De plus, les méthodes de sténotypie anglaises sont extrêmement précises, sans les confusions de sons que nous avons vues plus haut.

En conclusion, la transcription en français s'appuie bien sûr sur une recherche exhaustive dans un dictionnaire sténo-français, mais impose ensuite une stratégie de choix parmi le nombre élevé de possibilités. C'est ce que nous allons examiner dans la partie suivante.

3. CHOIX D'UN MODELE DE LANGAGE.

La recherche exhaustive dans le dictionnaire fournit un graphe de possibilités pour une phrase sténotypée donnée. Les noeuds de ce graphe sont les points entre les syllabes sténo, et les mots possibles sont des arcs reliant ces noeuds.



A priori, plusieurs stratégies sont envisageables pour effectuer un choix.

3.1. COINCIDENCE MAXIMALE.

Nous rappelons ici la méthode utilisée pour transcrire la sténotypie anglaise [36] et expliquée en I.2.3. Elle est satisfaisante en anglais, mais ne donne aucun résultat en français pour les raisons exposées précédemment. Elle donnerait par exemple: "Lait homme nouons regardé" pour "Les hommes nous ont regardé". En moyenne ce principe donne 50% d'erreurs pour le français...

3.2. GRAMMAIRE.

Les grammaires ont été souvent utilisées dans les traitements du langage naturel pour décrire les phrases permises, que ce soit avec le formalisme des ATN (Augmented Transition Network [42]), des réseaux procéduraux [38], ou des grammaires à stratégie programmable [34].
L'avantage de l'approche grammaticale est de pouvoir utiliser des contraintes globales au niveau de la phrase entière, et de garder un petit nombre de phrases candidates. Cependant ces modèles ne permettent de décrire qu'un sous-ensemble du langage naturel, plus ou moins grand suivant l'importance de la grammaire. Ceci entraîne deux sortes de problèmes:

- le peu de robustesse aux phrases prononcées incorrectes, qui ne peuvent être analysée par la grammaire. La cause peut être soit que la phrase est réellement incorrecte, soit qu'un mot est inconnu, soit qu'il y a une erreur de frappe dans

le cas de la sténotypie, soit enfin que certaines règles manquent à la grammaire. Si aucune phrase candidate n'a d'analyse, le système ne peut rien transcrire.

- les cas d'analyses multiples. Que choisir si plusieurs transcriptions ont une analyse?

En outre, plus on augmente le nombre de règles pour diminuer les cas du premier type, plus on risque d'avoir de cas du second type... Pour éviter cela, il faudrait prendre en compte des informations sémantiques, ce qui est faisable avec un petit vocabulaire, mais n'a pas été résolu pour un très large dictionnaire.

Remarque.

Matrices de précédence fréquentielles.

Andreewski et Fluhr [4] ont proposé un analyseur syntaxique utilisant des matrices de succession d'ordre deux ou trois. Une phrase est dite correcte si elle possède une chaîne de catégories grammaticales sous-jacente où toutes les successions de deux ou trois catégories sont corrects. L'apprentissage des matrices se fait semi-automatiquement ([13]). Les résultats en transcription phonèmes-graphèmes que nous en connaissons sont exposés dans [4]. Pour réduire le nombre de phrases candidates après une telle analyse, le système utilise des règles d'accord, et un critère de nombre minimum de mots. ([4]). Cependant, d'après notre expérience exposée dans la suite de ce travail, d'une part les fréquences lexicales nous paraissent absolument nécessaires pour obtenir une bonne performance, et d'autre part le critère de nombre minimum de mots est trop restrictif (voir II.3.2).

De plus un problème se pose là aussi lorsqu'aucune transcription ne possède de "chemin complet" binaire ou ternaire.

3.3. SOURCE DE MARKOV.

Des modèles probabilistes de langage issus de la Théorie de l'Information ont été introduits par Jelinek [6] pour la reconnaissance de parole. Comme nous reprendrons cette notion en détail dans la partie suivante, disons simplement qu'ils permettent d'attribuer une probabilité à toute suite de mots. La transcription choisie est alors celle de probabilité maximale. Ils s'appuient pour des raisons de calcul sur un contexte limité de chaque mot. Leurs avantages sont nombreux:

- Pas de situation où aucune transcription n'est possible.

- Pas de problèmes d'analyses multiples puisqu'elles sont ordonnées.

- L'apprentissage des probabilités élémentaires est automatique.

- Enfin, cette approche très souple permet de construire une large classe de modèles, dans laquelle on peut choisir en fonction de l'application.

Le contexte limité (généralement deux ou trois mots) fait la faiblesse de ces modèles, mais aussi leur robustesse: les erreurs ne se propagent pas à plus de trois ou quatre

mots. Les résultats de Jelinek en reconnaissance de parole pour 5.000 mots anglais sont excellents ([8]).

Pour notre application, avec un gros vocabulaire et surtout le discours oral libre, il nous a paru que l'approche grammaticale passerait avant tout par une étape manuelle d'écriture de règles, alors que les paramètres d'un modèle de Markov peuvent être "appris" de façon entièrement automatique dès que l'on possède suffisamment de textes français en machine. Plus séduisant encore, ces paramètres sont "optimaux" en regard des textes d'apprentissage. Le modèle est donc le meilleur possible en un sens qui sera précisé dans la partie suivante.

PARTIE II: CHOIX MARKOVIEN.

1. MODELE PROBABILISTE DE LANGAGE.

Pour la tâche de transcription qui était la nôtre, et plus généralement pour la reconnaissance de la parole avec large vocabulaire, nous avons vu qu'il est nécessaire d'avoir un modèle de langage qui participe au choix des candidats pour la suite de la phrase. L'idée de construire un modèle probabiliste adapté à cette tâche provient naturellement du fait que la succession des mots dans une phrase est soumise à des contraintes d'ordres grammatical et sémantique. Le principe est d'estimer la probabilité conditionnelle de l'apparition d'un mot, le début de la phrase étant fixé.

1.1. ETUDE INTUITIVE.

Soit un vocabulaire V, ensemble des mots $m_1, m_2 ... m_N$. Rigoureusement, la probabilité qu'un certain mot soit prononcé devrait dépendre de tous les mots l'ayant précédé. Si nous considérons comme événements les chaînes de mots, la probabilité d'une phrase $m(1,2...,n) = m_1 m_2 m_3 ... m_n$ s'écrirait comme le produit des probabilités conditionnelles:

$$P(m(1...n)) = P(m_1) \prod_{i=2}^{n} P(m_i / m(1,2,...,i-1))$$

Ces probabilités seraient

- difficiles à estimer, (à moins que ce ne soit à partir d'un corpus géant)
- et de toutes façons impossibles à stocker sur ordinateur.

1.1.1. Simplification.

Un moyen simple de réduire le nombre d'états (d'événements) à considérer est d'opérer une classification des chaînes de mots, ou encore de définir une équivalence entre elles. Par exemple,

- L'équivalence est donnée par les N derniers mots ("N-grammes"): deux chaînes $m_1m_2m_3...m_k$, $w_1w_2w_3...w_h$ sont déclarées équivalentes si les n derniers mots sont identiques dans chacune. Ainsi, seules les probabilités $p(m / m_{k-n}...m_k)$ doivent être estimées. En pratique, n=2 ou 3. La prédiction est faite de la façon suivante: supposons que l'on ait transcrit les mots "sur le ". Alors, pour la suite de la phrase, on choisira parmi les mots apparus souvent après ce couple dans le corpus d'apprentissage. On fera aisément la différence entre 'sur le bord' et 'sur le bol' par exemple, la fréquence du premier triplet étant bien supérieure à celle du second.
- L'équivalence entre chaînes de mots peut aussi utiliser une analyse syntaxique, ou plus simplement une caractérisation grammaticale des mots précédents: en français, après un couple (article adjectif) on prédira plus volontiers un substantif.

Il suffira d'estimer les probabilités $P(m / C_n)$ (où C_n est une des classes d'équivalence décrites) en relevant dans le corpus échantillon, c'est-à-dire l'ensemble des données

textuelles dont on dispose, les fréquences $f(m,C_n)$. Autrement dit:

$$P(m / C_n) = \frac{\text{nombre d'occurrences}(m,\text{chaines dans la classe } C_n)}{\text{nombre d'occurrences}(\text{chaines dans la classe } C_n)}$$

La structure probabiliste de l'ensemble (classes,mots,fréquences) est assez bien prise en compte, au point de vue théorie de l'information, par la notion de source de Markov, comme nous le verrons au §2 de cette partie.

1.1.2. Choix des classifications.

Nous venons de voir que plusieurs classifications en général sont disponibles. Il est clair qu'elles n'ont pas toutes les mêmes qualités:

- Si la classification est trop grossière, (mélange de mots très différents dans leur comportement), la prédiction sera mauvaise, mais le nombre de classes étant restreint, il suffit de peu de données.
- Ou bien, si elle trop fine, le nombre de classes est plus grand, il est difficile d'avoir assez de données pour obtenir des fréquences représentatives. Même s'il y avait assez de données, les fréquences à stocker seraient trop nombreuses aux points de vue volume et facilité d'accès.

Une solution possible (compromis) consiste à combiner linéairement les probabilités provenant de plusieurs classifications. Les poids de chacune doivent être choisis de façon "optimale" (Voir §II.1.3.2).

1.2. RAPPELS: THEORIE DE L'INFORMATION.

Pour la clarté de l'exposé, nous reprendrons dans cette section partie des outils et méthodes décrits dans [27].

1.2.1. Chaîne de Markov.

La notion de Chaîne de Markov est classique en théorie des probabilités. Rappelons qu'une chaîne de Markov d'ordre N est une structure où le passage par un état du système dépend de façon probabiliste des N états précédents. Deux états donnés sont reliés par au plus une transition, affectée d'une probabilité.

Reprenons l'exemple déjà cité (II.1.1) des 3-grammes. Nous associons au langage naturel une Chaîne de Markov, dont les états sont les couples de mots, les symboles sont les mots prononcés, les paramètres les fréquences relatives extraites du texte échantillon:

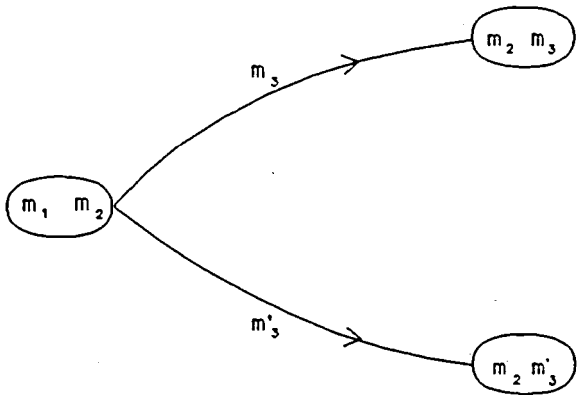


Figure 2: Chaîne de Markov associée au modèle des trigrammes (morceau)

Les probabilités de transition sont obtenues par:

$$f(m_3 / m_1,m_2) = \frac{\text{nombre d'occurrences de } (m_1,m_2,m_3)}{\text{nombre d'occurrences de } (m_1,m_2)}$$

Cette fréquence relative représente la probabilité d'apparition du mot m_3, sachant que les deux mots précédents étaient m_1,m_2.

1.2.2. Chaîne Cachée.

L'introduction de " Chaîne cachée " (Hidden Markov Chain), encore appelée Source de Markov, permet, lorsque le nombre d'états est très grand, de réduire la complexité du système, grâce au fait que deux états peuvent être reliés par plusieurs transitions. La figure 3 page 30 montre un exemple de source produisant des symboles pris dans (0,1). La définition précise est la suivante:

Une source de Markov est constituée,

- d'un ensemble fini d'états où sont spécifiés un état initial et un état final, s_I et s_F,
- d'un ensemble de symboles ou alphabet fini ,
- d'un ensemble de transitions. Chaque transition t "relie" un état de départ noté L(t) à un état but noté R(t) en produisant un symbole. (En particulier, une paire donnée d'états peut être reliée par plusieurs transitions). Les paramètres de la source sont les probabilités $q_s(t)$ pour tout état s et transition t (probabilité que t soit choisie partant de s).

Un "chemin" est une suite de transitions $t_1 t_2 ... t_n$, telles que l'état but de t_i soit le départ de t_{i+1}. Il est complet si la source de t_1 est l'état initial et le but de t_n est l'état final. Sa probabilité est simplement le produit des probabilités des transitions qui le composent. Une suite de symboles peut en général être produite par plusieurs chemins. Sa probabilité est alors la somme des probabilités des chemins qui la produisent.

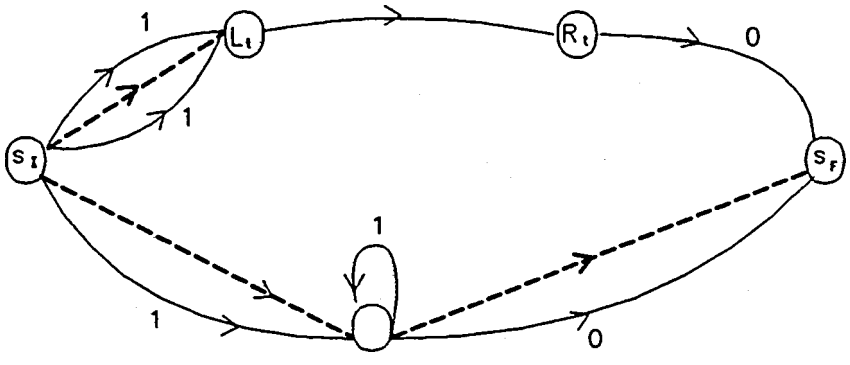


Figure 3: Exemple de source de Markov

Remarque: on peut autoriser des transitions ne produisant pas de symbole (flèches pointillées). Il suffit pour cela de rajouter un symbole (vide) à l'alphabet. Nous verrons au §II.1.3 l'utilité de ces transitions nulles lorsque l'on s'intéresse aux langages naturels.

1.2.3. Estimation des paramètres d'une source de Markov.

Nous voulons prendre une source de Markov pour modèle d'un phénomène réel (le langage). Pour cela, il faut que les paramètres de transitions soient adaptés aux données de la réalité. On dispose d'une longue séquence de symboles observés $X_1X_2...X_n$, (un long texte).

1.2.3.1. Idée intuitive.

Il s'agit de calculer les paramètres $q_s(t)$ de façon que la probabilité de la suite observée soit maximale. Si la suite est "très" longue, le nombre de fois où la transition t a été utilisée à partir de l'état s à un endroit de la production de $X_1X_2...X_n$, rapportée au nombre de passages par l'état s, est une bonne approximation de la probabilité $q_s(t)$ d'emprunter t partant de s. Bien entendu, les transitions ne sont pas observables directement, seule la suite l'est.

1.2.3.2. Algorithme de Baum.

Soit $P_i(t,X(1...n))$ la probabilité jointe que X(1...n) soit observée et que la transition t ait été utilisée pour produire X_i (ou juste avant la production de X_i si t est une

transition nulle). Le nombre de fois où t a été utilisée peut être estimé *a posteriori* par:

$$c(t,X(1...n)) = \sum_{i=1}^{n} P_i(t,X(1...n)) \,/P(X(1...n))$$

La fréquence relative d'utilisation de t à partir de l'état s se calcule à partir de ce compte:

$$f_s(t,X(1...n)) = \frac{c(t,X(1...n))}{\sum_{L(t')=s} c(t',X(1...n))}$$

si s=L(t) (départ de t). D'où la procédure itérative suivante, appelée algorithme de Forward Backward: ([27])

1. Choix de $q_s(t)$ initiaux arbitraires.
2. Calcul des $P_i(t,X(1...n))$ pour tous (i,t) à l'aide des $q_s(t)$.
3. Calcul des $f_s(t,X(1...n))$, ce qui donne de nouveaux $q'_s(t) = f_s(t,X(1...n))$
4. Poser $q_s(t) = q'_s(t)$ et reprendre à 2.

Pour calculer $P_i(t,X(1...n))$ en fonction des $q_s(t)$, on remarque que cette probabilité est le produit de trois facteurs:

- probabilité qu'une suite de transitions arrivant en L(t) produise X(1...i-1).
- probabilité que t soit prise partant de L(t), c'est-à-dire $q_s(t)$,
- probabilité qu'une suite de transitions partant de R(t) produise X(i+1,...n).

Notons $\alpha_i(s)$ la probabilité qu'une suite de transitions arrivant en s produise X(1...i), et $\beta_i(s)$ la probabilité qu'une suite de transitions partant de s produise X(i,...n). Alors,

$$P_i(t,X(1...n)) = \alpha_{i-1}(L(t)) \times q_s(t) \times \beta_{i+1}(R(t))$$

ou, si t est une transition nulle:

$$P_i(t,X(1...n)) = \alpha_{i-1}(L(t)) \times q_s(t) \times \beta_i(R(t))$$

Les $\alpha_i(s)$ s'obtiennent par récurrence ascendante sur i, les $\beta_i(s)$ par récurrence descendante sur i, d'où le nom d'algorithme de Forward Backward donné par Jelinek à cette procédure.

1.2.3.3. Solution mathématique.

La probabilité de la suite observée est une fonction polynomiale F de plusieurs variables et homogène des $q_s(t)$. L.E Baum a démontré la convergence du procédé vers une solution, maximum local de cette fonction ([9]). La preuve s'appuie sur les propriétés de la transformation qui, à une famille de paramètres $q_s(t)$, associe la famille de nouveaux paramètres $q'_s(t)$ calculés au pas 3 de l'algorithme précédent. Essentiellement,

$$F(q_s(t)) \geq F(q'_s(t))$$

avec égalité si et seulement si les $q_s(t)$ sont un extremum local de F.

I.3. APPLICATION AU LANGAGE NATUREL.

La source de Markov choisie pour modéliser le langage naturel dépend bien sûr de la taille du vocabulaire et de l'application. Cependant, il sera difficile de trouver une classification unique des chaînes de mots qui soit satisfaisante. La qualité de la prédiction dépend étroitement du corpus où sont relevées les fréquences, comme le montrera au §II.5 l'étude des résultats de transcription avec différentes tailles de corpus. On peut déja citer un exemple de cette dépendance que nous avons expérimentée lors de l'élaboration d'un modèle de langage pour la transcription de la sténotypie: nous avions d'abord relevé les fréquences de triplets de classes grammaticales sur 65.000 mots provenant essentiellement du livre "Histoire de l'informatique", de René Moreau. La transcription, s'appuyant sur ces statistiques, d'un discours oral de De Gaulle a montré des fautes caractéristiques sur les tournures interrogatives ("quand est il des obstacles...") fréquentes à l'oral mais rares dans le type de discours écrit considéré pour l'apprentissage.

Jelinek rapporte dans [27] l'expérience suivante: les fréquences de trigrammes sur 1.000 mots de vocabulaire ont été relevées sur 1,5 million de mots. Le modèle est testé sur un nouveau texte de 300.000 mots: environ 23% des trigrammes du nouveau texte n'ont jamais été observés dans le corpus d'1,5 million et ont donc une probabilité nulle.

1.3.1. Solution possible.

On a vu (§ II.1.1.2) les avantages respectifs de plusieurs classifications $C_1,C_2,...,C_m$ de plus en plus grossières. Combinons les fréquences $f(m_n / C_i(m_1...m_{n-1}))$ linéairement pour déclarer que la probabilité résultante d'un mot m_n sachant $m_1,m_2,...,m_{n-1}$ s'écrit

$$P(m_n / m_1...m_{n-1}) = \sum_{i=1}^{m} \lambda_i(N_1^{n-1}) \times f(m_n / C_i(m_1 m_2...m_{n-1}))$$

les λ_i sont positifs ou nuls, de somme égale à 1. On choisit les poids λ_i fonction de N_1^{n-1}, soit $(N_1,N_2,...,N_{n-1})$ où N_i est le nombre de fois où la classe $C_i(m_1...m_{n-1})$ à été observée dans le corpus. En choisissant les λ de cette manière, on s'assure que même si les données sont insuffisantes pour représenter la classification fine, la fréquence des classes plus grossières (qui, elles, sont représentées) compense la valeur de la probabilité.

Par exemple, on peut compenser le manque de trigrammes (m_1,m_2,m_3) cité au début de ce paragraphe en combinant les fréquences des trigrammes avec celles des bigrammes (m_2,m_3). Intuitivement, si (m_1,m_2). =("programme prévu") n'a pas eu d'occurrences ni par conséquent aucun 3-gramme commençant par ces deux mots dans le corpus échantillon, et que "programme prévu" se présente dans un test, on a peut-être rencontré le mot "prévu" seul pour choisir quel mot devrait venir ensuite. La formule ci-dessus s'écrit simplement

$$P(m_3 / m_1 m_2) = \lambda \times f(m_3 / m_1,m_2) + (1 - \lambda) \times f(m_3 / m_2)$$

Dans le cas où le couple (m_1,m_2) a été recontré zéro fois, $f(m_3 / m_1,m_2)$ ne serait pas même définie, cependant $P(m_3 / m_1 m_2)$ a un sens. Le poids peut donc être choisi pour chaque terme (3- et 2-gramme) en fonction du nombre d'occurrences du couple (m_1,m_2):

Si ce nombre est élevé, alors le premier terme de la somme est fiable, et λ doit être proche de 1.

Sinon, c'est $1-\lambda$ qui doit être plus grand et le deuxième terme contribuer à P.

1.3.2. Calcul des poids.

Retournons au problème général concernant un nombre quelconque de classifications. A chaque classification, on peut associer une source de Markov, dont les états sont les classes, les symboles les mots.

On peut également voir la combinaison de plusieurs de ces modèles comme une nouvelle source de Markov, obtenue en adjoignant des transitions nulles (ne produisant pas de symbole) et des états supplémentaires, représentant les classes d'équivalence pour chaque état, comme le montre la figure suivante:

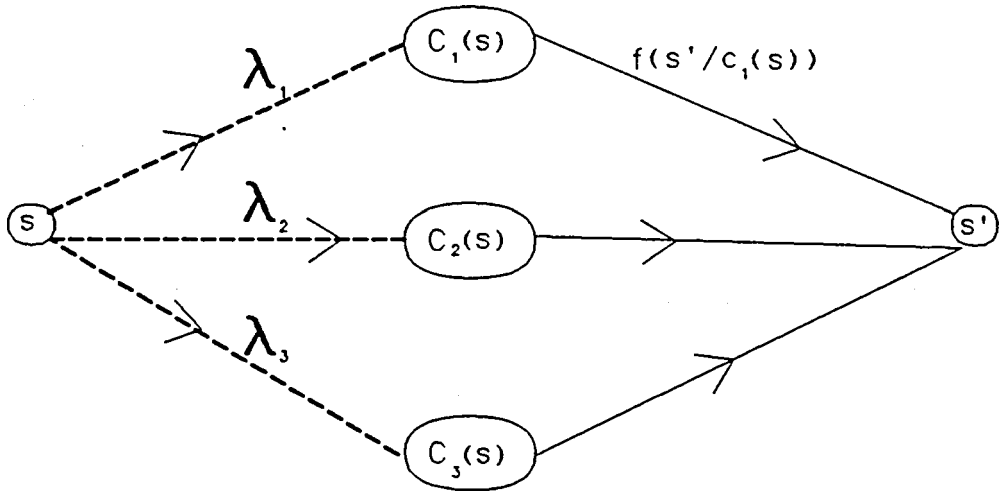


Figure 4: Source de Markov obtenue par interpolation

La probabilité d'aller dans l'état s' à partir de s est alors bien la somme des probabilités des chemins de transitions qui y conduisent, c'est-à-dire la combinaison linéaire entre les fréquences. Les λ sont alors des paramètres particuliers de cette source, (les probabilités des transitions nulles), calculables comme nous l'avons vu au §II.1.2.3.

Cependant, si l'on applique le procédé tel quel, les poids obtenus sont adaptés à la réalité du corpus échantillon. Or la classification la plus fine suffit à prédire une phrase faisant partie du corpus. On obtient donc un poids égal à 1 pour celle-ci, et 0 pour les autres... Pour éviter cette difficulté, Jelinek [27] propose les variantes suivantes de l'algorithme de Forward Backward:

Soit (Deleted Estimator), séparer les données en n blocs. Pour chaque i entre 1 et n, on conduit le Forward-Backward comme en 4.2.3 de manière à maximiser la probabilité du ième bloc, en utilisant les fréquences relevées sur la réunion des

(n-1) autres blocs. Ces fréquences sont d'ailleurs laissées fixes, seuls les λ sont calculés itérativement. On somme ensuite et on normalise les λ trouvés avec chaque bloc.

Soit (Held-out Estimator) [27], séparer les données en deux parties: l'une servant de base pour la collecte des fréquences, l'autre jouant le role de nouveau texte pour ajuster les poids.

1.3.3. Avantages.

Cette représentation probabiliste est intéressante à plusieurs points de vue,

- Il est possible d'estimer automatiquement les paramètres du modèle.
- La combinaison de plusieurs modèles de ce genre (correspondant à différentes classifications) permet de se restreindre à un corpus échantillon d'une taille raisonnable.
- Une phrase entière est considérée comme un chemin complet (au sens de I.1.2.2) de la source.

  Sa recherche se fait par des méthodes classiques, par exemple l'algorithme de Vitterbi, quand le nombre d'états n'est pas trop grand:

  Connaissant une suite de symboles $y_1...y_n$ produits par une source, il s'agit de trouver le chemin complet $t_1...t_n$ de probabilité maximale parmi ceux qui produisent $y_1...y_n$.

  - Pour k compris entre 1 et n, et s un noeud du graphe (un état de la source), soit $T_k(s)$ le chemin le plus probable arrivant en s et produisant $y_1...y_k$, soit $P_k(s)$ sa probabilité.

- Pour trouver $T_n(s_f)$ (s_f étant l'état final), la méthode consiste à calculer par récurrence sur k les $T_k(s)$ et $P_k(s)$ pour tout s.

$T_{k+1}(s)$ et $P_{k+1}(s)$ sont obtenus en prenant le maximum, sur les transitions t arrivant en s et produisant y_{k+1}, des produits $P_k(s') \times q_{s'}(t)$. avec s' départ de t. En pratique, on ne retient que les états s pour lesquels $P_k(s)$ est supérieur à un certain seuil.

1.4. EXEMPLE DETAILLE.

A titre d'exemple, nous détaillons ici un travail effectué en collaboration avec F. Jelinek pour la reconnaissance de parole en anglais ([26]). La tâche considérée est la dictée automatique en mots isolés. Le vocabulaire est tiré d'un large échantillon de correspondance (1 million de mots) et comprend les 5.000 mots les plus courants. Le but du modèle qui va suivre est d'améliorer les performances du modèle basé sur une interpolation linéaire entre les fréquences relatives (trigrammes, bigrammes, unigrammes). Avec cette interpolation, le taux d'erreur (reconnaissance en mots isolés) est de 5,2%. Les erreurs sont réparties également entre la comparaison acoustique et le modèle de langage.

Parmi les erreurs provenant du modèle de langage, on trouve des exemples comme "Part time of test are you going" au lieu de "What kind of test are you running", qui suggèrent que la classe grammaticale du mot à prédire pourrait être utilisée avec profit. Même si la probabilité acoustique de "running" était bonne, le trigramme "are you going" était beaucoup plus fréquent que "are you running" dans le corpus d'apprentissage. L'idée est d'utiliser le fait que dans les mots fréquents après "are you", on trouve beaucoup de formes en "ing", (même si ce n'est pas "running"),

et donc toute forme de ce type devrait avoir une assez bonne probabilité dans cette situation.

1.4.1. Définition de la nouvelle source.

Soient g_1,g_2 des classes grammaticales, m_1,m_2 des mots.

Soit $h(g / m_1,m_2)$ la probabilité d'avoir à la suite de m_1,m_2 une classe g, et $k(m_3 / g)$ la probabilité de choisir m_3, sachant que la classe est g. La source de Markov associée peut se représenter:

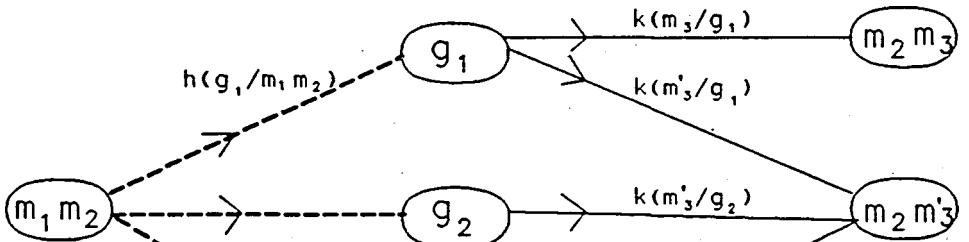


Figure 5: Modèle avec utilisation des classes grammaticales

Un estimé Pr_1 de la probabilité de m_3 sachant (m_1,m_2) est donné par:

$$Pr_1(m_3 / m_1,m_2) = \sum h(g / m_1,m_2) \times k(m_3 / g)$$

La somme est faite sur toutes les classes possibles du mot m_3.

De même on peut estimer:

$$Pr_2(m_3 / m_2) = \sum h'(g / m_2) \times k(m_3 / g)$$

en prédisant la classe à partir d'un seul mot précédent.

1.4.2. Choix des classes.

Nous avons choisi:

- 27 classes grammaticales: nom, adjectif, verbe etc.,
- 11 classes à "contenu sémantique": prix, pays, nom de personne, ville, date, etc.

En fait ces classes concernent des mots ou groupes de mots. Dans le texte d'apprentissage on a remplacé par exemple les dates par un seul mot " date". Ainsi, le modèle prédira une date (toutes confondues pour les calculs), un sous modèle permettra ensuite de choisir une date particulière. L'avantage ici est la flexibilité du vocabulaire. On pourrait aussi choisir d'autres classes sémantiques, telles que verbes d'action, nom abstrait, etc.

1.4.3. Méthode.

Il s'agit de calculer les distributions h,k,h'.

1.4.3.1. Initialisation

- Chaque mot du vocabulaire est associé à ses classes possibles (d'après le dictionnaire).
- Les $k'(g / m)$ (probabilités d'une classe sachant le mot) sont initialement équivalentes, égales à: 1/nombre de classes pour ce mot.
- Les $k(m / g)$ sont égaux à $k'(g / m)f(m) / f(g)$ (formule de Bayes). Les f(g) sont simplement obtenues comme somme sur m des $k'(g / m)f(m)$.
- Les $h(g / m_1,m_2)$ sont obtenus par comptage pour tous g, m_1,m_2 :

$$h(g / m_1,m_2) = (1 / N(m_1,m_2)) \times \sum_{m_3} N(m_1,m_2,m_3)k'(g / m_3)$$

N désigne le nombre d'occurrences.

1.4.3.2. Ajustement.

La méthode de L.E Baum (§II.1.2.3) s'applique pour ajuster les probabilités de transition de la source de Markov représentée dans la figure 5 page 40, c'est-à-dire les $h(g / m_1,m_2)$ et $k(g / m_3)$. La distribution $h'(g / m_2)$ est calculée en sommant sur m_1 les $h(g / m_1,m_2)$.

1.4.3.3. Observations.

Les distributions h et k obtenues reflètent la répartition des classes sur un même mot, par exemple pour le prénom Ken, existant aussi comme verbe rarement employé, on trouve une probabilité 0.95 d'être un nom propre, 0.05 d'être un verbe, cela malgré

le fait que les probabilités initiales étaient égales. Cette convergence se produit au bout de trois itérations seulement. Il est clair que ces nombres se rapprocheraient encore de 1 et 0 respectivement si l'on effectuait un plus nombre d'itérations. Elles montrent aussi où le choix de certaines classes est inadéquat: pour les formes en "ed" comme opened etc., le partage entre verbe et adjectif s'est fait généralement largement au profit de la classe adjectif. Cela s'explique par la taille de cette classe comparée à celle des verbes, et suggère qu'une nouvelle classe prétérit serait utile.

1.4.4. Modèle final.

Un premier estimé était déjà obtenu par combinaison des fréquences relatives

$$Proba_1(m_3 / m_1,m_2) = \lambda_1 \times f(m_3 / m_1,m_2) + \lambda_2 \times f(m_3 / m_2) + \lambda_3 \times f(m_3).$$

Un deuxième estimé est la combinaison de Pr_1,Pr_2:

$$Proba_2(m_3 / m_1,m_2) = \lambda'_1 \times Pr_1(m_3 / m_1,m_2) + \lambda'_2 \times Pr_2(m_3 / m_2)$$

Le modèle final combine ces deux prédictions:

$$Proba(m_3 / m_1 m_2) = \Gamma_1 \times Proba_1(m_3 / m_1,m_2) + \Gamma_2 \times Proba_2(m_3 / m_1,m_2)$$

Il est illustré dans la figure suivante:

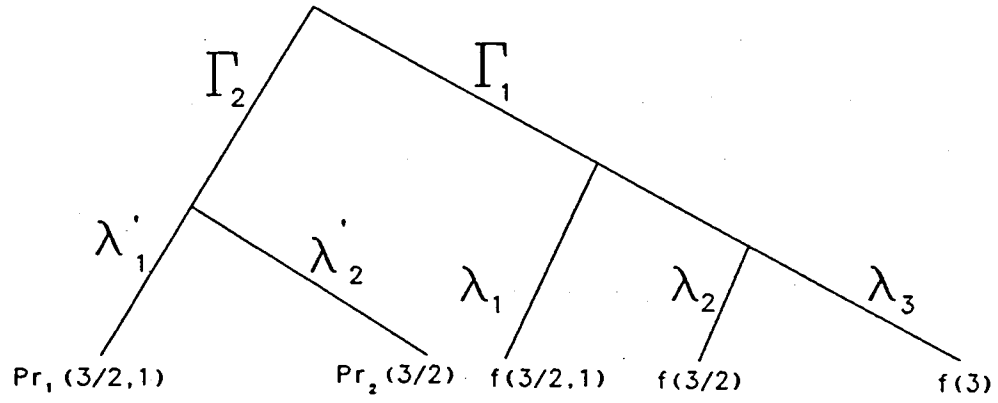


Figure 6: Combinaison finale des fréquences et des prédictions par classes

1, 2, 3 = trois mots.

f = fréquences relatives

1.4.4.1. Calcul des coefficients.

Le corpus est séparé en deux parties:

Texte1, 80% des données, avec lequel les coefficients inférieurs λ,λ' vont être calculés.

Texte2, 20% des données, avec lequel les coefficients Γ vont être calculés.

- $\lambda_1,\lambda_2,\lambda_3$ sont ajustés par un deleted estimator sur Texte1 (II.1.3.2), avec quatre blocs. Détaillons par exemple le calcul de λ_1 avec un des quatre blocs. Soit t_1 la transition nulle de probabilité λ_1, de départ (m_1,m_2) et de but $R(t_1)$ (c'est l'état $c_1(s)$ sur la figure 4 page 37). m_3 est le ième mot de la suite observée. Reprenant les expressions de II.1.2.3,

$$P_i(t_1,Bloc) = Proba_1(->m_2) \times \lambda_1 \times f(m_3 / m_1,m_2) \times Proba_1(m_3->)$$

où $Proba_1(->m_2)$ désigne la probabilité que la suite jusqu'à m_2 soit produite, et $Proba_1(m_3->)$ désigne la probabilité que la suite à partir de m_3 soit produite. Les fréquences sont relevées sur les trois autres blocs. $Proba_1(Bloc)$ est le produit de ces deux derniers facteurs et de:

$$Proba_1(m_3 / m_1m_2) = \lambda_1 \times f(m_3 / m_1,m_2) + \lambda_2 \times f(m_3 / m_2) + \lambda_3 \times f(m_3).$$

On peut donc simplifier l'expression

$$\frac{P_i(t_1,Bloc)}{Proba_1(bloc)} = \frac{\lambda_1 \times f(m_3 / m_1,m_2)}{\lambda_1 \times f(m_3 / m_1,m_2) + \lambda_2 \times f(m_3 / m_2) + \lambda_3 \times f(m_3)}$$

La transformation à itérer est finalement:

$$\lambda_1 = \sum_{Bloc} \frac{\lambda_1 \times f(m_3 / m_1,m_2)}{\lambda_1 \times f(m_3 / m_1,m_2) + \lambda_2 \times f(m_3 / m_2) + \lambda_3 \times f(m_3)}$$

Deux formules analogues donnent λ_2,λ_3, qui seront ensuite normalisés.

- λ'_1,λ'_2, sont ajustés de la même façon, avec un deleted estimator sur Texte1. Les calculs précédents sont valables en remplaçant les fréquences relatives par Pr_1 et Pr_2.

- Γ_1,Γ_2, sont obtenus sur Texte2, avec cette fois un Held out Estimator (II.1.3.2). On itère la transformation:

$$\Gamma_1 = \sum_{Texte2} \frac{\Gamma_1 \times Proba_1(m_3 / m_1,m_2)}{\Gamma_1 \times Proba_1(m_3 / m_1,m_2) + \Gamma_2 \times Proba_2(m_3 / m_1,m_2)}$$

où $Proba_1$ et $Proba_2$ sont obtenues sur Texte1.

1.4.5. Résultats.

L'inconvénient d'une prédiction basée uniquement sur les fréquences (trigrammes, bigrammes, unigrammes) est que dans un grand nombre de cas la probabilité du troisième mot en fonction des deux précédents est très faible($<10^{-4}$). Le nouveau modèle permet de "rattraper" un certain nombre de ces mots.

Une récente expérience de reconnaissance sur quatre locuteurs dictant le même ensemble de phrases, soit 300 mots, montre la répartition suivante (avec le modèle 1, c'est-à-dire combinaison des fréquences):

En tout, 63 erreurs ont été faites, soit un taux de 5,2% d'erreurs (sur les 1200 mots).

Sur 48 cas où la probabilité linguistique était inférieure à 10^{-5} , 14 erreurs ont été faites (soit 30%).

Sur 24 cas où la probabilité linguistique était entre 10^{-5} et 10^{-4}, 5 erreurs ont été faites (soit 22%).

Comparaison entre modèle1 (fréquences) et modèle final sur le même texte de 300 mots:

Modèle 1	Modèle final
12 proba $< 10^{-5}$	2 proba $< 10^{-5}$
6 proba entre 10^{-5} et 10^{-4}	11 proba entre 10^{-5} et 10^{-4}

Globalement pour les quatre locuteurs, parmi les 19 mots ayant provoqué des erreurs avec le modèle 1, 16 ont avec le modèle final une probabilité au moins dix fois supérieure.

Les 13 trigrammes à risque d'erreurs avec le modèle final (probabilité $\leq 10^{-4}$) sont tous inclus dans les 18 du modèle 1. Il n'y a donc pas de nouveau mot à haut risque d'erreur avec le modèle final. Bien sûr, la somme des probabilités étant toujours 1, si certaines augmentent, d'autres diminuent. En fait, augmenter la probabilité de la suite de mots correcte doit diminuer celle des suites incorrectes, donc le taux de reconnaissance doit augmenter.

En conclusion, le nombre global de mots à haut risque d'erreur à cause d'une probabilité linguistique trop faible a diminué de 1/3. Il est donc raisonnable d'espérer que cette méthode diminuera le nombre d'erreurs dues au modèle linguistique.

2. CHOIX D'UN MODELE POUR LA TRANSCRIPTION STENOTYPIE-FRANCAIS.

2.1. DEFINITION.

Si dans le système de reconnaissance de 5.000 mots, il est possible de prendre pour états d'une source de Markov les couples de mots, cela est tout à fait exclu dans le cas de 150.000 mots. Le nombre d'états possibles serait énorme. Les paramètres seraient impossibles à stocker, même si l'on disposait d'un corpus géant pour les calculer.

Par contre, si l'on choisit comme états les couples de classes grammaticales, le nombre d'états est très réduit. Notre système de classes ayant 92 éléments, une telle source aura de toute façon moins de 10.000 états.

D'autre part, les chemins de cette source doivent être des phrases. Les symboles produits par la source seront donc les mots. Entre deux états (p_1,p_2) (p_2,p_3) il y aura autant de transitions possibles que de mots m_3 de la classe p_3, qui chacune émettent comme symbole précisément le mot m_3.

Les 92 classes que nous avons décrites dans la partie I et mises dans le dictionnaire

ont été choisies pour être bien adaptées à la prédiction à partir de deux mots (voir Annexe 1):

- les pronoms personnels sujets prédisent un verbe conjugué à la même personne, d'où la séparation entre les pronoms personnels et les autres, et la marque de personne pour les premiers. Les autres pronoms objets (me, te, se) n'impliquent pas une personne particulière, ils ne portent donc pas cette marque. Un mot comme "nous" est bien sûr affecté des deux classes de pronoms. L'idée ici est que le couple "Pronom personnel 3ème personne, verbe 3ème personne" aura une probabilité beaucoup plus forte que le couple "Pronom personnel 3ème personne, verbe 2ème personne", d'où la possibilité de choisir correctement "il chantait" plutôt que "il chantais". On voit ici un intérêt du modèle: l'accord doit se faire automatiquement grâce au choix des classes et aux probabilités. En fait, certaines des triclasses obtenues comme très probables, seront reconnues après coup équivalentes à des règles d'accord simple. Nous vérifierons cela au § II.2.2 en étudiant la liste des triclasses fréquentes.
- L'accord entre déterminants, substantifs et adjectifs (du moins lorsqu'ils sont proches) se fera grâce aux mêmes principes: il suffit de créer quatre classes de substantifs (SUBSMS, SUBSFS, SUBSMP, SUBSFP), d'adjectifs (ADJEMS, ADJEFS, ADJEMP, ADJEFP), et de déterminants (DETRMS, DETRFS, DETRMP, DETRFP).
- Les verbes conjugués portent une marque de personne, ce qui donne 6 classes: VERB1, VERB2, VERB3, VERB4, VERB5, VERB6, car d'une part les différentes inflexions de "personne" sont souvent homonymes, et d'autre part elles peuvent être prédites par le contexte proche, comme on l'a vu par exemple plus haut pour les pronoms personnels. Par contre, les temps et mode ne pourront être prédits

avec un contexte limité à deux mots. Il était donc inutile de les indiquer dans les classes.

- Les pronoms relatifs forment quatre classes PRELMS, PRELFS, PRELMP, PRELFP. Cela permet essentiellement la transmission du genre et du nombre de l'antécédent dans un cas comme:

  "la femme qui est venue"

  Si le relatif ne portait pas de genre, "venue" n'aurait pas de raison d'être préféré à "venu", puisque la prédiction se ferait à partir de PREL AUXE3 seulement. Au contraire, du fait que le mot "qui" est doté potentiellement des quatre classes, on peut espérer que le modèle prédise "qui PRELFS" à partir de DETRFS SUBSFS, et ensuite "venue PPASFS" à partir de PRELFS AUXE3.

Nous avons maintenant une structure possible de modèle, qui peut se représenter par la figure suivante:

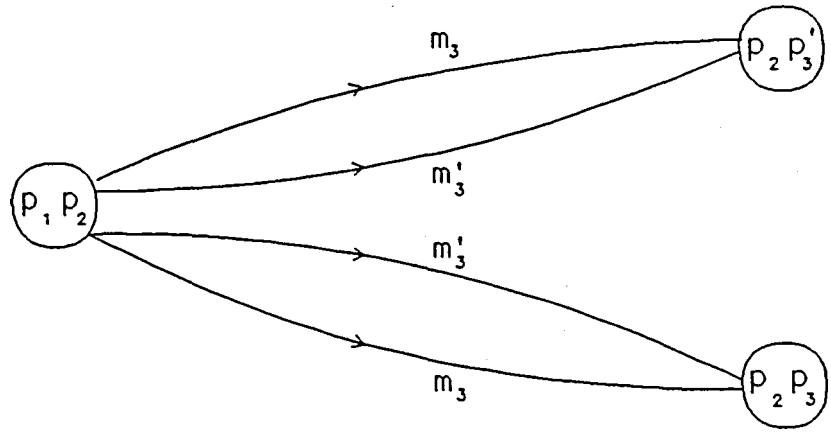


Figure 7: Fragment du modèle triclasse

Une transition est donnée par un triplet de classes (p_1,p_2,p_3) et un mot de classe p_3. Nous appellerons désormais biclasse le couple (p_1,p_2) et triclasse le triplet (p_1,p_2,p_3).

Si nous adoptons cette source telle quelle, le problème des données insuffisantes cité au §II.1.3 se posera: si une triclasse (p_1,p_2,p_3). n'est jamais apparue dans le corpus d'apprentissage, elle risque tout de même d'apparaître dans un texte à décoder, et p_3 ne pourra être prédite à partir de (p_1,p_2). C'est pourquoi nous avons en fait élaboré un modèle interpolé comme il est expliqué au §II.1.3: dans un cas de triclasse (p_1,p_2,p_3) manquante, utiliser la prédiction basée sur la biclasse (p_2,p_3) qui, elle, a peut-être été vue dans le corpus. La source de Markov correspondante au modèle interpolé possède des états supplémentaires fictifs, et des transitions nulles. Elle est représentée par la figure suivante:

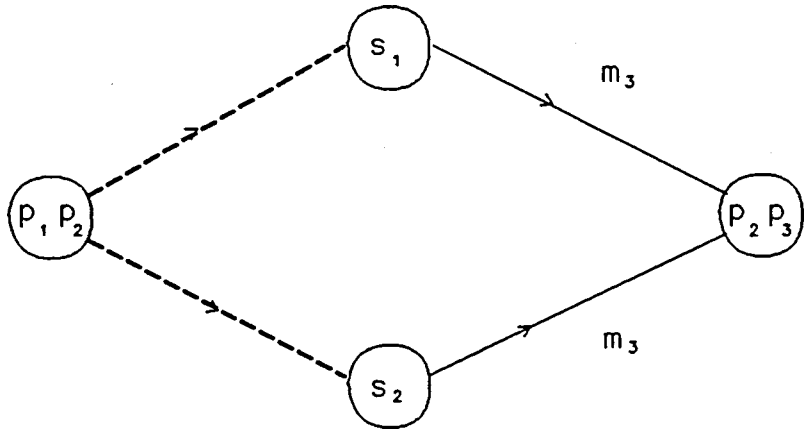


Figure 8: Modèle interpolé biclasses-triclasses

L'état fictif s_1 symbolise la biclasse (p_1,p_2) tandis que l'état fictif s_2 symbolise la classe d'équivalence de toutes les biclasses ayant le même p_2.

Comme au §II.1.3.2, les probabilités des transitions nulles (flèches pointillées) vers chacun des deux états fictifs, sont des poids λ_1, λ_2, de somme 1, à ajuster de façon optimale pour combiner les fréquences relatives de $(m_3 / p_1,p_2)$ et de (m_3 / p_2). En fait, ce sont les fréquences de triplets de classes $(p_3 / p_1,p_2)$ et de couples (p_3 / p_2) que l'on va collecter et combiner. On approximera la probabilité $p(m_3 / p_1,p_2)$ par le produit de la probabilité que p_3 suive p_1,p_2, avec la probabilité de produire le mot m_3, sachant que la classe est p_3. Cette dernière distribution $k(m / p)$ est calculée simplement par le rapport du nombre d'occurences du mot m avec la classe p, et du compte total de la classe p. La probabilité de produire le mot m_3 à partir de l'état (p_1,p_2), est alors estimée par la somme des probabilités des deux chemins possibles:

$$p(m_3 / p_1,p_2) = (\lambda_1 f(p_3 / p_1,p_2) + \lambda_2 f(p_3 / p_2)) \times k(m_3 / p_3) + \varepsilon$$

Nous avons rajouté un ε (égal à 10^{-4}) afin de ne pas risquer d'obtenir 0 pour une triclasse, auquel cas on ne pourrait prédire aucune phrase la contenant. Nous allons maintenant voir comment calculer ces différents paramètres.

2.2. ETIQUETAGE.

L'étiquetage a pour but de collecter les fréquences de base des triclasses et biclasses, dans le corpus d'apprentissage. La fréquence relative d'apparition d'une classe p_3

sachant les deux classes précédentes s'écrit comme le rapport des comptes:

$$f(p_3 / p_1,p_2) = \frac{\text{nombre d'occurrences de } (p_1,p_2,p_3)}{\text{nombre d'occurrences de } (p_1,p_2)}$$

De même,

$$f(p_3 / p_2) = \frac{\text{nombre d'occurrences } (p_2,p_3)}{\text{nombre d'occurrences } (p_2)}$$

Le corpus d'apprentissage a été constitué des textes français dont nous disposions au Centre Scientifique sous forme lisible par machine. On y trouve aussi bien de l'oral (conférences d'hommes politiques), que de l'écrit assez libre (journaux divers) ou plus fermé (la Constitution). On trouvera en "Annexe J. Liste des textes d'apprentissage" page 155 la liste exacte de ces textes. Cette grande variété fait que le modèle ne sera adapté à aucun domaine ou locuteur particulier, mais représentera plutôt une moyenne. Il est clair que cela est justifié par la tâche particulière de transcription de la sténotypie. Par définition, elle est utilisée dans toutes sortes de conférences, discours, débats, etc. Pour une application plus spécifique avec un domaine de discours ou un thème bien cerné, il y aurait intérêt à sélectionner des textes du domaine.

Pour pouvoir calculer les fréquences, il faut que ces textes aient été préalablement "étiquetés": autrement dit que chaque mot ait été associé à sa classe grammaticale dans son contexte. Pour ce faire, nous avons suivi la procédure semi-automatique suivante:

- Sur 2.000 mots, nous avons marqué la classe manuellement et collecté les comptes de biclasses résultant.
- Avec le modèle de Markov basé sur ce petit nombre de biclasses, nous avons étiqueté automatiquement 16.000 nouveaux mots, en choisissant pour chaque phrase la suite de classes la plus probable, par un algorithme semblable à l'alignement de Viterbi (§II.1.3.3). Les comptes de biclasses et triclasses sur les 18.000 mots étiquetés ont été relevés.
- Par interpolation entre ces statistiques triclasses et biclasses (avec des poids arbitraires), un modèle un peu meilleur est obtenu et utilisé pour étiqueter automatiquement 47.000 nouveaux mots. Les erreurs d'étiquetage sur ces 47.000 mots (erreurs de classes dans le dictionnaire, confusion entre participes passés en "é" et verbes à cause de l'absence d'accents dans certains textes...) ont été corrigées à la main (3 jours). A nouveau, les triclasses et biclasses sont recueillies sur l'ensemble des 65.000 mots.
- Le reste du corpus qui comprend au total 1,2 million de mots a été étiqueté automatiquement avec les statistiques de l'étape précédente, cette fois sans correction manuelle. Une évaluation partielle a montré un taux d'erreur inférieur à 5%.

Les annexes B et C montrent les 100 biclasses et triclasses les plus fréquentes trouvées dans le corpus total, par ordre de fréquence décroissante (pour un total de 4.400 biclasses différentes et 50.000 triclasses). Beaucoup d'entre elles sont reconnaissables comme des structures syntaxiques simples. Regardons d'abord les biclasses:
On trouve en haut de la liste le groupe nominal DET N (déterminant nom) pour les quatre couples (genre, nombre) possibles: DETfs Nfs, DETms Nms, DETmp Nmp, DETfp Nfp, ainsi que le groupe nominal non déterminé N A (quatre biclasses).
Se trouvent également les compléments:

(à, de) (Nms, Nfs, Nmp, Nom propre)

(au, du) Nms

des (Nmp, Nfp)

Il manque (à,de) Nfp dans les 100 premières biclasses, mais après vérification cette biclasse est tout de même fréquente, c'est la 117eme de la liste. On peut noter cependant la fréquence significativement plus grande de "à" ou "de" suivi d'un nom singulier (environ 10.000) que de "à" ou "de" suivi d'un nom pluriel (2.500). Ceci est dû sans doute aux expressions courantes comme "homme de loi", "temps de guerre" etc. qui, si elles ne sont pas toutes figées, sont des tournures extrêmement productives.

On trouve enfin les structures de phrases:

il verbe 3ème personne singulier

je verbe 1ère personne singulier

L'absence des structures analogues aux autres personnes montre d'ailleurs combien le corpus privilégie certains styles de discours: à la première personne les discours d'hommes politiques (on aurait alors pu attendre aussi la première personne du pluriel: la biclasse "nous, verbe, 1ère personne pluriel" est en effet en position 142, donc assez fréquente), ou à la troisième dans les textes écrits.

Les biclasses dont le premier élément est le point nous renseignent sur les classes en début de phrase, essentiellement:

DETs, "il", PREP, Conjonction de coordination, adverbe, "c'", "je".

L'observation des 100 triclasses les plus fréquentes amène des remarques analogues sur des structures à trois éléments:

Le groupe nominal DET N A (déterminant nom adjectif) est présent aux quatre catégories de genre et nombre.

Le groupe DET A N est seulement au singulier (masculin ou féminin).

On observe N (à de) N fréquent avec les deux noms au singulier. Il s'agit sûrement de N de N en fait, mais les prépositions "à" et "de" sont dans la même classe. Ceci s'accorde avec la remarque faite plus haut concernant les biclasses (à, de) suivi d'un nom singulier. (Nms, Nfs) du Nms, et Nfs des (Nmp, Nfp) sont très fréquentes, bien qu'un peu moins que les précédentes. On reconnaît également les compléments suivants:

(à, de) DET N
PREP DET N
des Np Ap
du Nms Ams

Enfin nous trouvons des structures de phrase:

DETs Ns verb 3ème personne singulier
Detfs Nfs auxe 3ème personne singulier

Toutes les biclasses ou triclasses fréquentes ne recouvrent pas des structures, comme (Nms (à, de)) ou (DETfs Nfs conjonction de coordination) par exemple. Cependant il était intéressant de vérifier que les règles d'accord simple, local, (et seulement

celles là bien sûr) s'y trouvent toutes représentées, grâce à quoi le modèle markovien va pouvoir les contenir.

Les deux figures suivantes montrent le nombre de triclasses et de biclasses différentes rencontrées en fonction de la taille du corpus. Avec 1,2 million de mots, nous avons rencontré 4400 biclasses, soit environ 50% des couples *a priori* possibles. La courbe a une pente assez faible à 1,2 million, ce qui laisse penser qu'au moins les statistiques de biclasses sont représentatives et raisonnablement fiables.
Par contre, seulement 50.000 triclasses ont été en tout rencontrées, soit 5% des triplets possibles. La courbe cette fois montre que le corpus d'un million de mots est encore incomplet pour les triclasses.

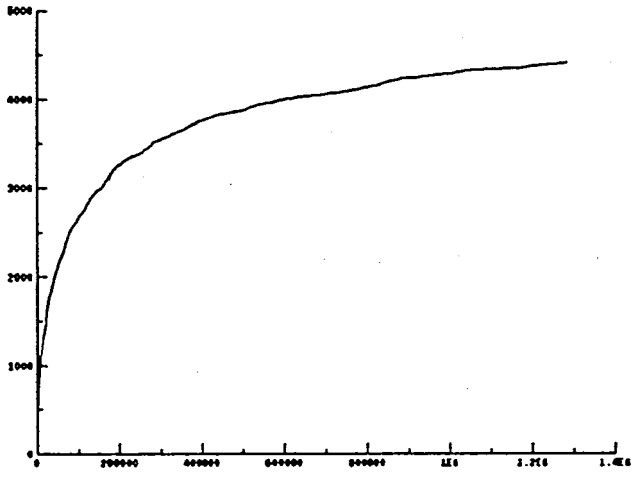


Figure 9: Nombre de biclasses rencontrées en fonction du nombre de mots dans le corpus d'apprentissage.

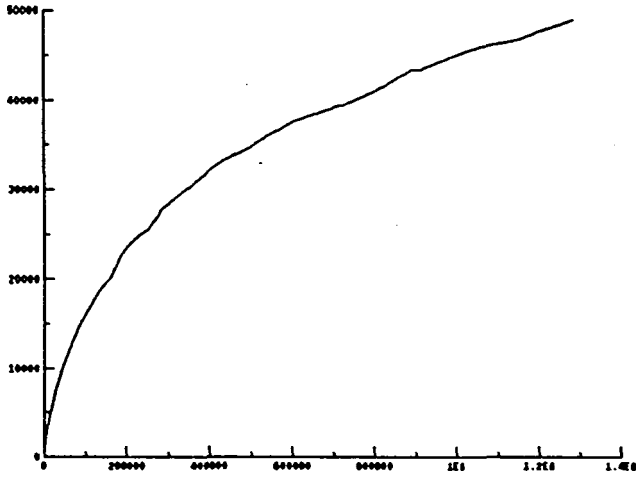


Figure 10: Nombre de triclasses rencontrées en fonction du nombre de mots dans le corpus d'apprentissage.

Ces observations ne font que confirmer l'intérêt que présente une interpolation linéaire entre les deux distributions. Voyons maintenant le calcul des coefficients de l'interpolation.

2.3. ESTIMATION DES POIDS.

Les poids λ_1, λ_2, dépendent de (p_1,p_2) puisqu'il y a deux états fictifs ajoutés à la source pour chaque état (p_1,p_2) (§II.2.1, figure 8 page 52). Cependant, plutôt que de calculer autant de poids que d'états, on peut contraindre les λ dans l'algorithme de Forward Backward (II.1.2) de plusieurs manières:

1. soit, comme dans Jelinek [27], ils sont une fonction (affine par intervalles) du compte de la biclasse (p_1,p_2), avec la justification donnée en général au II.1.3.1.

2. soit, on les prend fonction uniquement de la dernière classe (p_2). Intuitivement, certaines classes prédisent bien la suivante sans besoin de l'avant-dernière: par exemple les articles (qui prédisent fortement substantif ou adjectif) ou les prépositions. D'autres comme la négation, les adverbes, ne suffisent pas à prédire la classe suivante ("je ne veux" et "il ne veut").

Nous avons conduit le Forward Backward (II.1.2.3) des deux manières, avec différentes tailles de corpus d'apprentissage: 65.000, 200.000, et 1 million de mots. Pour la première version, 10 intervalles ont été déterminés pour les comptes de biclasses, et un couple (λ_1,λ_2), calculé par intervalle. Pour la seconde, il y a un couple pour chacune des 92 classes grammaticales. La comparaison des résultats de transcription pour ces 6 ensembles de coefficients est donnée plus loin. (§II.5).
Le corpus est découpé en deux parties (Held Out estimator) 75% et 25%. Les fréquences de triclasses et biclasses sont collectées sur 75%, tandis que les poids sont ajustés sur les 25% restant qui jouent le rôle de nouveau texte. L'algorithme se ramène à itérer la transformation:

$$\lambda_1 = \sum \frac{\lambda_1 f(p_3 / p_1,p_2)}{\lambda_1 f(p_3 / p_1,p_2) + \lambda_2 f(p_3 / p_2)}$$

$$\lambda_2 = \sum \frac{\lambda_2 f(p_3 / p_2)}{\lambda_1 f(p_3 / p_1,p_2) + \lambda_2 f(p_3 / p_2)}$$

La somme est faite sur les triplets dans les 25%. Les valeurs des fréquences proviennent des 75%. L'algorithme converge très rapidement (4 itérations). Les deux figures suivantes montrent les valeurs de λ_1 obtenues avec le corpus total.

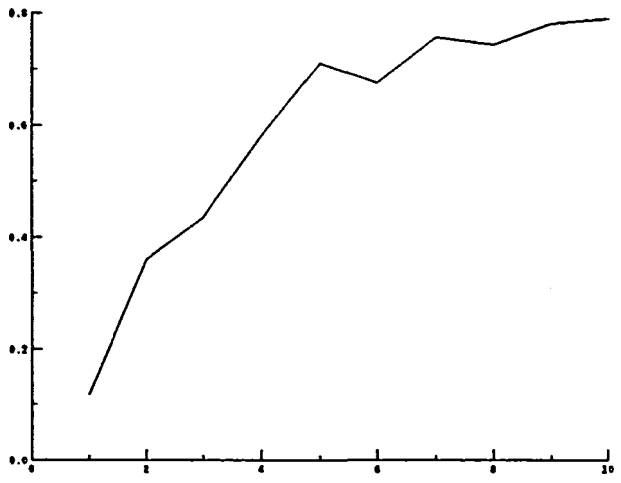


Figure 11: Poids fonction du compte de la biclasse

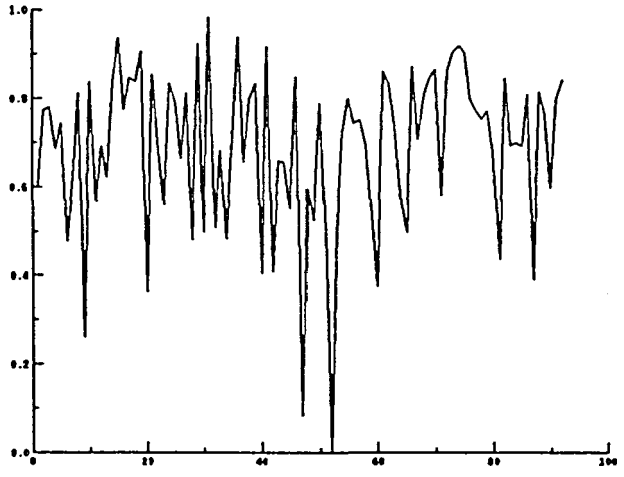


Figure 12: Poids fonction de la dernière classe

Pour les poids dépendant de p_2, on trouve des valeurs de λ_1 grandes pour la négation (classe 36) et l'adverbe (classe 10), et faibles pour la préposition (classe 80), ce qui confirme l'intuition de départ. Le point aberrant de la courbe où l'un des coefficients est nul correspond en fait à une classe jamais rencontrée dans le corpus

(pronom indéfini féminin pluriel).

Pour les valeurs calculées à partir des comptes de biclasses, λ_1 est une fonction à peu près croissante du compte de (p_1,p_2), comme on s'y attendait.

3. DECODAGE

Les paramètres du modèle de langage étant calculés, nous décrivons maintenant comment ce modèle est utilisé dans le décodage proprement dit de la bande sténo. Le principe comme nous l'avons énoncé à la fin de la partie I est de choisir la transcription de probabilité maximale, la probabilité d'une suite de mots $m_1 m_2 ... m_p$ s'écrivant dans notre modèle:

$$P(m_1 ... m_p) = \prod_{n-3}^{p} p(p_n / p_{n-2} p_{n-1}) k(m_n / p_n)$$

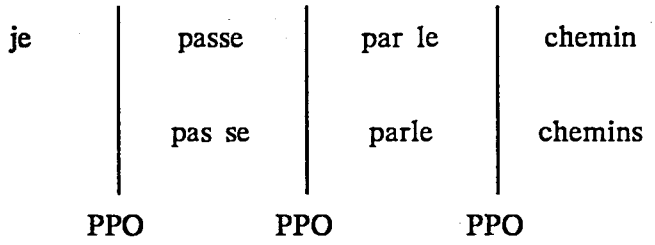
où $p(p_n / p_{n-2} p_{n-1})$ est obtenue par l'interpolation décrite plus haut.

Cependant nous avons souligné dans la partie I la croissance exponentielle du nombre de transcriptions en fonction du nombre de mots de la phrase. Il est donc hors de question d'énumérer toutes les suites de mots possibles entre deux signes "*" de sténo, "*" étant le symbole du point. On adopte au contraire une stratégie sous-optimale, qui ne produit pas forcément "la" solution la plus probable. L'algorithme de "stack decoding" mis au point par Jelinek pour la reconnaissance de la parole était aussi de cette sorte. Il compare des chemins de plus en plus longs en ne gardant que ceux dont la fonction de probabilité est supérieure à un certain seuil. Cette fonction de probabilité est préférée à la probabilité elle-même d'une part pour qu'elle ne décroisse pas quand la longueur du chemin augmente, et d'autre part pour rendre

comparables des chemins de longueurs différentes ([27]). Dans la transcription de la sténotypie, nous avons pu profiter des particularités de cette tâche pour mettre au point un algorithme plus rapide qui permet de ne pas attendre la fin de la phrase avant de fixer la transcription du début, et de comparer toujours des chemins de même longueur sténo. Il s'appuie sur l'existence de "points de passage obligé" que nous allons expliciter dans ce qui suit.

3.1. POINTS DE PASSAGE OBLIGE.

A partir de la bande sténotypée, une recherche dans le dictionnaire sténo-français permet d'obtenir tous les mots dont la sténotypie figure sur la bande, ainsi que leurs classes respectives. On obtient un graphe dont les noeuds sont les séparations entre syllabes sténotypées et les arcs sont les mots, chaque arc indiquant les syllabes utilisées par le mot associé (voir figure 2 page 17). Dans ce graphe, certains noeuds servent pour tous les chemins complets possibles reliant la première "*" (début de phrase) à la suivante (fin de phrase). On les appelle "points de passage obligé" (PPO). En particulier, ce sont toujours des frontières de mot.



Supposons que l'on soit situé à un PPO. Pour déterminer le suivant, on cherche dans le dictionnaire tous les mots commençant à ce noeud et coïncidant avec une partie de la sténo qui vient ensuite. On itère cette recherche en commençant maintenant plus à droite, à partir de la première frontière des mots trouvés. Lorsqu'on arrive à un point qui est la frontière la plus à droite de tous les mots trouvés jusque là, on a affaire à un PPO. On peut remarquer que ce procédé, tel qu'il est, ne donne pas forcément le premier PPO à droite, à cause des "branches mortes" du graphe, c'est-à-dire des arcs qui ne possèdent pas de prolongement jusqu'à la fin de la phrase. L'exemple suivant illustre cette situation:
La sténo est "NOU-A-FON-PRI-A-KT-*", pour le français "nous avons pris acte". A cause du mot "pria" le procédé ne trouvera pas de PPO entre les syllabes PRI et A, bien que tous les chemins allant jusqu'au bout de la phrase passent par ce noeud, puisqu'il n'existe aucun mot ayant la sténo "KT".
Expérimentalement, on trouve un PPO tous les deux ou trois mots, bien qu'il soit théoriquement possible de ne pas en rencontrer, dans une phrase "pathologique" du genre "coucou coucou..." répété un grand nombre de fois...Cette propriété vérifiée expérimentalement fait penser à la propriété de synchronisation des codes ([37]), bien que la sténo ne soit pas un code au sens algébrique (si elle l'était, il y aurait justement une transcription unique!).

Entre deux PPO, on a un ensemble de chemins possibles qui correspondent chacun à un ou plusieurs mots. Le graphe peut donc être "factorisé" comme une suite d'ensembles de groupes de mots:

$$G = eg_1.eg_2.eg_3....eg_n$$

Choisir une phrase du graphe revient à choisir un groupe dans chaque ensemble. Le nombre total de phrases possibles est:

$$N(G) = N(eg_1).N(eg_2).....N(eg_n)$$

On voit maintenant l'intérêt des points de passage obligé: à chaque étape de l'énumération, les chemins correspondent à la même portion de la sténotypie, et donc leurs probabilités sont comparables. On ne compare jamais de chemins de longueurs (au sens du codage) différentes, et on évite ainsi les problèmes de normalisation de la probabilité en fonction de la longueur qui se posent dans le cas général.

3.2. PREFILTRE.

Le décodage s'effectue phrase par phrase. On opère de gauche à droite, de PPO en PPO.

Supposons avoir deux points de passage obligés, PPO_1 et PPO_2 tels que:

- la transcription a été fixée jusqu'à PPO_1.
- entre PPO_1 et PPO_2 nous avons p chemins candidats d'au moins deux mots chacun.

On détermine alors le prochain PPO, soit PPO_3, tel que tout chemin candidat entre PPO_2 et PPO_3 contienne au moins 2 mots. PPO_3 est le premier ou le deuxième PPO trouvé par le procédé précédent. On effectue un préfiltre sur tous les chemins entre PPO_2 et PPO_3 consistant à:

1. garder seulement les chemins dont le nombre de mots est minimum, minimum plus 1, ou minimum plus 2. Par exemple, si le mot "entendu" est possible, on gardera les transcriptions de 1, 2 et 3 mots, comme "entendu, en tendu, an tant dû" etc. On éliminera "en t'en dus, an t'an t'eu" etc...

   Remarque: on pourrait à première vue penser qu'il n'est pas nécessaire de garder les chemins avec le nombre minimum de mots plus deux. C'est d'ailleurs l'hypothèse que nous avions faite au début de ce travail... Mais un certain nombre de cas comme "genevois, je ne vois", "je t'envoie la, jetant voilà", "sévère, c'est vers" etc. ont conduit à adopter le critère final.

2. calculer pour chaque chemin une probabilité d'après un modèle simplement basé sur les biclasses, et garder les dix meilleurs parmi ceux dont la probabilité est supérieure à 10^{-3} fois la plus grande.

A la sortie de ce préfiltre, il reste q chemins sélectionnés entre PPO_2 et PPO_3 (avec $q \leq 10$). Expérimentalement, ce préfiltre n'a jamais éliminé à tort le bon mot. D'autre part, le facteur de branchement, c'est-à- dire le nombre moyen de mots possibles à chaque syllabe sténo, est passé de 10 avant le préfiltre (§I.2) à 1,4 à la sortie. C'est dire l'énorme avantage de cette présélection qui réduit la complexité de façon appréciable sans faire rien perdre en qualité.

3.3. CHOIX.

Pour choisir définitivement la transcription entre PPO_1 et PPO_2 on calcule les probabilités des $p \times q$ concaténations d'un chemin entre PPO_1 et PPO_2, et d'un chemin entre PPO_2 et PPO_3. Ces probabilités font appel cette fois au modèle de Markov interpolé avec biclasses et triclasses. Le chemin entre PPO_1 et PPO_2, apparaissant dans la meilleure concaténation est alors fixé comme transcription. On recommence la même opération en repartant de PPO_2: recherche de PPO_4, préfiltre entre PPO_3 et PPO_4, choix entre PPO_2 et PPO_3, etc. jusqu'à la fin du texte.

Voici un exemple de déroulement du décodage pour le début de phrase "n'oublions pas que le premier ...". On trouvera l'exemple complet en "Annexe F. Trace du décodage d'une phrase" page 143. Chaque mot est suivi de sa classe grammaticale et de sa fréquence dans cette classe. Chaque groupe de mots dans le préfiltre est précédé de sa probabilité biclasse.

```
***** PRE-FILTRE *****

  1 4.0670E-10(nous PPER4 1.00E+00)(plions VERB4 1.28E-04)
  2 6.2630E-11(ne NE     1.00E+00)(oublions VERB4 3.08E-03)
  3 6.0846E-13(nous PPOBMP 1.71E-01)(plions VERB4 1.28E-04)
***** CHOIX *****

 1/. .
***** PRE-FILTRE *****

  1 4.2901E-14(pas PAS    5.63E-01)(que CSUB  5.53E-01)(le DETRMS
              3.82E-01) (premier ADJEMS 8.47E-03)
  2 2.1107E-15(pas PAS    5.63E-01)(que CSUB  5.53E-01)(le PPOBMS
              5.54E-01)(premier ADJEMS 8.47E-03)
  3 1.9762E-16(bâcles VERB2 3.80E-06)(premiers ADJEMP 7.33E-03)
  4 1.6596E-16(bâcles VERB2 3.80E-06)(premier ADJEMS 8.47E-03)
  5 1.3221E-16(pas PAS    5.63E-01)(que CSUB  5.53E-01)(le DETRMS
              3.82E-01)(premiers ADJEMP 7.33E-03)
```

```
  6 1.1155E-16(bâcle VERB1 8.77E-06)(premier ADJEMS 8.47E-03)
  7 7.4288E-17(bâcle VERB3 4.13E-06)(premier ADJEMS 8.47E-03)
  8 7.2264E-17(bas ADVE  3.31E-04)(que CSUB  5.53E-01)(le DETRMS
3.82E-01)(premier ADJEMS 8.47E-03)
  9 4.8872E-17(bâclent VERB6 5.70E-06)(premiers ADJEMP 7.33E-03)
***** CHOIX *****

 2/ne oublions
***** PRE-FILTRE *****

  1 4.4897E-11(a AUXA3 5.42E-01)(été AUXE  3.07E-01)
  2 4.3159E-11(à PDEA  3.14E-01)(été AUXE  3.07E-01)
  3 3.8964E-12(à PDEA  3.14E-01)(aider VINF  1.46E-03)
  4 2.5011E-12(à PDEA  3.14E-01)(été SUBSMS 1.35E-03)
  5 3.3899E-13(a AUXA3 5.42E-01)(aidé PPASMS 6.73E-04)
  6 1.7790E-13(a AUXA3 5.42E-01)(était AUXE3 2.40E-01)
  7 1.5175E-13(ah ADVE  6.36E-03)(été AUXE  3.07E-01)
  8 1.1372E-13(à PDEA  3.14E-01)(était AUXE3 2.40E-01)
  9 1.1069E-13(à PDEA  3.14E-01)(étés SUBSFP2 .02E-04)
 10 1.1016E-13(ah ADVE  6.36E-03)(était AUXE3 2.40E-01)
***** CHOIX *****

 1/pas que le premier
```

4. IMPLEMENTATION.

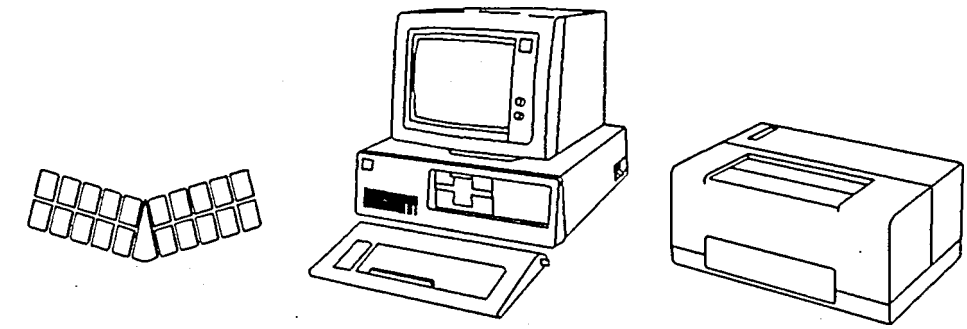
Le système de transcription dont nous venons de décrire les composantes principales:

- dictionnaire,
- modèle de langage,
- algorithme de décodage,

a été développé en PL1 sous VM sur l'IBM 3031 du Centre Scientifique. Il a été testé sous VM sur des données simulées, à savoir du texte sténotypé par programme. Dans un deuxième temps, le système, écrit cette fois en Pascal pour la plus grande part, a été entièrement transféré sur un micro ordinateur IBM PC-XT. Seules les routines d'accès au dictionnaire et aux statistiques du modèle de langage ont été mises en Assembleur (B.Merialdo) pour l'accès direct et parce que le Pascal ne pouvait pas gérer plus de 64K de variables. L'IBM PC est configuré avec un disque dur de 10 Mégaoctets où sont stockés les statistiques (400K) et le dictionnaire (2 Mega), et une mémoire vive de 512K.

Le clavier sténo étant entièrement mécanique, des contacts électriques ont été rajoutés à chaque touche. Ce clavier est relié directement à l'ordinateur, grâce à une carte prototype de décodage d'adresse branchée sur le bus de l'IBM PC (B. Denoix).

Le schéma général du processus de transcription est représenté sur la figure suivante:



La vitesse de transcription est de 100 mots par minute sur IBM PC-XT, et de 180 mots par minute sur IBM PC-AT, c'est-à-dire le temps réel (aussi rapide que le discours).

5. RESULTATS.

Sous VM, nous avons comparé l'influence de la taille du corpus d'apprentissage, et l'influence du Forward Backward, sur les résultats. Aussi nous avons mené une expérience de transcription pour chaque taille: 65.000, 200.000, 1 million de mots, et chaque ensemble de coefficients pour l'interpolation biclasse-triclasse: coefficients arbitrairement fixés à $\lambda_1 = 1$, $\lambda_2 = 0.01$, (bien que la somme ne soit pas 1, la transcription est la même que si l'on normalisait), ou bien dépendants de la dernière classe, ou bien dépendants du compte de la biclasse (§II.2.3). Dans tous les cas, le texte à transcrire comprenait 2800 mots, et provenait de journaux télévisés n'ayant pas fait partie du corpus d'apprentissage. On a compté comme erreurs les mots du texte de départ ne se retrouvant pas identiques dans la transcription. Les résultats sont résumés dans le tableau suivant:

taille corpus	coefficients arbitraires	coefficients fonction de p_2	coefficients fonction du compte
65.000 mots	379	302	302
200.000 mots	300	257	254
1,2 Mmots	263	240	247

Nombre de mots incorrects sur 2800

La taille de l'apprentissage a une très grande influence, ce qui n'est pas surprenant. Le passage de 65.000 à 1,2 million supprime entre 18% (avec Forward Backward) et 30% (sans) des erreurs.

L'utilisation d'un algorithme d'optimisation des poids supprime entre 8 et 20% des erreurs par rapport à des poids arbitraires. Par contre, les deux sortes de contrainte sur les poids sont équivalentes. De plus, les fautes de transcription sont à peu près les mêmes.

L'influence du Forward Backward se fait moins sentir quand la taille du corpus augmente. En effet, le nombre et la fiabilité des triclasses augmentent, et la combinaison avec les biclasses fait moins de différence.

Voici des exemples où l'on peut observer la qualité de la prédiction avec des corpus différents. Les mots en gras séparés par des slash indiquent les transcriptions différentes choisies respectivement par les modèles appris sur 65.000, 200.000 et

1,2 Million de mots, (avec des poids fonction de la classe précédente):

- "*Pour connaître les premières réactions à Cracovie* ***noua fond appeler / nous avons appelé / nous avons appelé*** *le correspondant...*"
- "*C'est lui que vous allez entendre vous* ***pourrait / pourrez / pourrez*** *voir...*"
- "*tous ses écrits* ***doutes / toutes / toutes*** *ses prises de position...*"
- "*La colonie polonaise de Paris* ***faite / faite / fête*** *à sa manière l'élection...*"
- "***Ce son airs qui on déclenchait / sont airs qui ont déclenché / sont elles qui ont déclenché*** *les hostilités...*"
- "*Mr de Guiringaud* ***à tous de mêmes ajouter / à tout de même ajouter / a tout de même ajouté***"

Une prédiction basée uniquement sur le contexte grammatical local est donc capable d'une performance de 91,3%. L'analyse des erreurs montre la répartition suivante:

- 1,7% de noms propres inconnus (particulièrement fréquents dans les journaux télévisés)
- 1% de frontières de mots incorrectes ("effet" au lieu de "et fait")
- 2,4% de fautes grammaticales
- 3,6% d'autres fautes.

Le problème des noms propres a été partiellement résolu dans le système mis au point sur IBM PC par la possibilité d'y placer à l'avance un dictionnaire propre à l'utilisateur, contenant les principaux noms de personne devant intervenir. De toutes façons, le système, grâce à la robustesse du modèle et à la propriété de synchronisation (§II.3.1), peut se "rattraper" au plus deux mots plus loin, ainsi les mots inconnus

ne propagent pas des erreurs dans toute la phrase.

Les accords grammaticaux sont impossibles à prédire par un tel modèle lorsque les mots en rapport sont distants dans la phrase, par exemple sujet et verbe séparés par une relative, un complément, etc. D'autres fautes grammaticales sont incluses dans les 2,4%, par exemple des confusions entre et/est, a/à, ou/où, etc. Nous y reviendrons dans la partie suivante.

Enfin, le dernier type d'erreurs relève de la sémantique, ou même de la pragmatique, très difficiles à traiter dans un cas de discours oral ouvert. Par exemple, "*qu'en pense ton arôme*" au lieu de "*qu'en pense t'on à Rome*" ... Dans le paragraphe II.6 suivant, nous donnons une étude partielle de ces fautes et des cooccurrences sémantiques qui pourrait réduire leur nombre.

Sur IBM PC, nous avons pu tester le système en situation réelle avec une sténo entrée directement au clavier. Plusieurs heures de discours ont ainsi été saisies puis transcrites. Un extrait de la transcription "brute" obtenue sur IBM PC à partir d'une saisie en situation réelle est donnée dans "Annexe G. Transcription d'un discours saisi en situation réelle." page 147. Nous avons constaté que le taux d'erreurs du système lui-même reste sensiblement 8%. A cela s'ajoutent bien sûr les fautes de frappe. Celles ci dépendent des sténotypistes, leur proportion varie entre 2 et 5%. Le pourcentage dépend aussi du type de discours, (en particulier de la vitesse d'élocution), et du clavier utilisé. Nous avons essayé plusieurs prototypes de sténotype électronique fournis par la société Grandjean. Toute faute de frappe sténo produira au moins une faute à la transcription. Soit la frappe fausse correspond à un mot dans le dictionnaire, mais erroné dans le texte, soit elle ne correspond à aucun mot existant, et elle est alors reproduite telle quelle dans le texte transcrit. Pour que le modèle de langage puisse continuer à travailler avec une suite de classes,

une frappe inconnue reçoit par défaut la classe grammaticale nom propre. Là encore, le contexte limité de la prédiction fait que les erreurs ne se propagent pas sur plus de deux mots, et n'interrompent pas le processus de transcription.

5.1. APPLICATIONS.

Le système portable sur micro ordinateur pourrait faire gagner un temps appréciable dans la transcription du discours sténotypé. Le temps de frappe dactylographiée se trouve remplacé par le temps de correction.

De plus le temps réel sur l'IBM PC-AT permet d'envisager une application plus particulière: le sous-titrage en direct des émissions de télévision. En effet la préparation de sous-titres est actuellement un travail très long: il faut entre 50 et 100 heures de travail pour sous-titrer une émission d'une heure. Ce temps est occupé par la saisie du script, le découpage en sous-titres, puis la synchronisation de ceux-ci sur le film. Cela limite le nombre d'heures qui peuvent être sous-titrées, malgré la forte demande de la part des malentendants, dont c'est le seul moyen d'accès à la télévision. En outre, il est impossible de sous-titrer des émissions en direct avec ces moyens.

Le procédé Antiope (développé par le CCETT), permet de superposer du texte à l'image et de réserver la réception du texte aux spectateurs possédant un décodeur. On pourrait donc envisager de sténotyper le texte de l'émission au fur et à mesure, de transcrire en français et d'envoyer la transcription brute comme sous-titre. Le texte ne serait décalé que de 2 à 4 mots par rapport à l'image. Ceci conviendrait très bien pour les journaux par exemple (voir [18]). Le CCETT a d'ailleurs commencé

une étude sur ce sujet ([23]). En Angleterre et aux Etats-Unis, où les systèmes de transcription de la sténotypie dont nous avons déjà parlé, sont disponibles, des expériences de sous-titrage en direct ont déjà été réalisées ([5], ([10]). Il faudrait bien sûr vérifier dans des conditions réelles si le taux d'erreurs dans la transcription, et surtout le genre d'erreurs, ne gênent pas la compréhension du texte.

6. ETUDE DES HOMONYMIES SEMANTIQUES.

6.1. OBSERVATIONS.

Parmi les 3,6% de fautes répertoriées comme "sémantiques", un certain nombre résultent d'un mauvais choix entre deux "homonymes", définis comme au §I.2.3.1, c'est-à-dire deux mots ayant la même sténo et une classe grammaticale en commun. Dans le modèle que nous venons de décrire, la différence de probabilité entre ces deux mots est due uniquement à leurs fréquences respectives dans la classe en question. Le premier terme de la formule du §II.2.1 (interpolation entre fréquences de la triclasse et de la biclasse) est en effet identique pour les deux mots. Le modèle choisit alors systématiquement le plus fréquent des deux. Ce genre de situation représente 1% dans le pourcentage total d'erreurs. Les exemples suivants sont tirés du texte donné en "Annexe D. Transcription d'un journal télévisé." page 135.

- "*cent quarante **maîtres***"
- "*l'expédition dont faisait **parti** Gérard...*"
- "*l'âge de son **chaise***"
- "*en se **passant** sur le fait*"

- "*je forme des feux pour le succès...*"

Il est assez remarquable qu'une grande proportion de ces erreurs se produise en fait dans des expressions courantes, (se baser sur, au second plan, former des voeux), ou dans des expressions figées (faire partie) où les éléments ne sont pas sémantiquement indépendants.

Les formes du dictionnaire qui possèdent au moins un homonyme, comprennent un grand nombre de verbes conjugués à la première ou troisième personne (ferai/ferais, meurt/meure). Si l'on exclut ces cas, il reste environ 7.500 mots concernés. Parmi ceux là, les plus difficiles à désambiguïser sont bien sûr les paires de mots dont chaque élément est fréquent. Pour les autres, on se trompera très rarement en choisissant systématiquement le plus fréquent. Nous avons donc retenu un ensemble de 300 paires fréquentes dont le début est donné en "Annexe H. Liste des paires les plus fréquentes" page 149 par ordre décroissant de fréquence du moins fréquent des deux...[3] En haut de cette liste, on ne sera pas surpris de trouver les couples:

- fois / voix / voie / foi
- mère / mer,
- coup / goût / cou / coût,
- plan / blanc,
- temps / dents etc.

---

3 Ces paires sont constituées de mots de même classe grammaticale, en particulier mêmes genre et nombre, sauf pour les substantifs pluriel où le genre peut être différent, car le contexte ne permet pas toujours de choisir le genre ("des temps"/"des dents").

6.2. METHODE.

Nous avons cherché une distribution qui remplace la fréquence du mot dans l'expression de sa probabilité et tienne compte du contexte de ce mot. Par contexte, nous entendrons les mots pleins (substantifs, adjectifs, verbes) rencontrés à droite et à gauche du mot. Pour des raisons de calcul nous limitons le nombre de ces mots du contexte à 10 de chaque côté. (Ils peuvent bien sûr être situés plus loin que 10 mots).

Soit m le mot à prédire à partir du contexte ctx. La probabilité conditionnelle d'apparition de m sachant le contexte peut s'écrire:

$$p(m / ctx) = \frac{p(ctx / m)p(m)}{p(ctx)}$$

Le dénominateur ne dépendant pas de m, nous prendrons en fait l'expression au numérateur, qui est une distribution sur les mots m à un facteur constant près. $p(m)$ est simplement la fréquence brute du mot trouvée dans le dictionnaire.

Le contexte ctx est formé des mots $m_1,m_2,...,m_n$. L'hypothèse selon laquelle les événements m_i / m, c'est à dire "m_i apparaît dans le contexte de m", seraient indépendants, n'est certainement pas parfaite puisque, si les apparitions d'un mot m_i et de m sont liées, celles de m_i et m_j ensemble le sont certainement. Cependant si l'on fait cette hypothèse, on trouve:

$$p(ctx / m) = \prod_{i=1}^{n} p(m_i \text{ en place i dans le contexte de m})$$

L'apprentissage devrait donc collecter pour chaque mot étudié (ambigu) tous les

mots apparaissant dans son contexte, leur position, leur nombre de cooccurrences, soit 20 comptes à tenir pour chaque mot. Cela serait énorme à stocker. Nous avons pris l'hypothèse simplificatrice que la distribution des cooccurrences à l'intérieur de la fenêtre de 20 mots est uniforme, autrement dit que la position n'a pas d'importance. On peut voir dans "Annexe I. Cooccurrences dans le contexte de mère" page 153 où les mots trouvés dans le contexte de "mère" ont été marqués avec le nombre d'occurrences dans chaque quart de fenêtre, de la position -10 à -6, puis -5 à -1, 1 à 5, 6 à 10, que c'est souvent vérifié, sauf dans des cas d'expression courante comme "mère de famille", ou "son père et sa mère". Il suffit alors de collecter le nombre de contextes du mot m où le mot m_i apparaît, soit $n(m_i,m)$. Si $n(m)$ désigne le nombre total de fois où m apparaît dans le corpus, $p(m_i$ en place i dans le contexte de $m)$ est alors approximé par :

$$\frac{1}{20} \times \frac{n(m_i,m)}{n(m)}$$

Encore une fois la constante 20 au dénominateur ne dépend pas de m et nous ne la gardons pas pour le calcul. L'expression finale de la distribution (à un facteur constant près) est donc:

$$p(m \,/\, ctx) = p(m) \times \prod_{i=1}^{20} \frac{n(m_i,m)}{n(m)}$$

Cependant, si l'on calcule l'espérance de la variable: nombre de cooccurrences entre un mot quelconque apparaissant p fois et un mot donné apparaissant q fois dans un corpus de taille T, on trouve cqp/T, (où c est la taille de contexte choisie), lorsque l'on suppose tous les mots équirépartis ([31]). Dans notre cas, c=20, et T=650.000

est le nombre de mots pour l'apprentissage. Par exemple le mot "mère" y apparaît 131 fois. Pour un mot quelconque de fréquence 50, l'espérance de le rencontrer dans le contexte de "mère" serait donc 0,2 s'ils étaient indépendants. La probabilité de le rencontrer:

0 fois est à peu près	0,8
1 fois	0,2
3 fois ou plus	0

On peut donc considérer que les cooccurrences trouvées égales à 1 ou 2 ne sont pas très significatives, puisqu'elles ont de bonnes chances de se produire même lorsque les mots sont équidistribués. Aussi nous ne tiendrons compte que de celles supérieures ou égales à 3. Dans la formule ci-dessus donnant $p(m / ctx)$, nous remplacerons le terme

$$\frac{n(m_i,m)}{n(m)}$$

par ε fixé à 2.10^{-4} , dans les cas où $n(m_i,m)$ est inférieur à 3. Pour une justification du choix de cette valeur, on peut se reporter à [31].

6.3. RESULTATS.

Dans les 650.000 mots de l'apprentissage, nous avons pu recueillir des données contextuelles sur 190 des 300 homonymes fréquents. La distribution basée sur ces cooccurences a été testée sur un autre corpus de 580.000 mots. 6.074 mots ambigus

(appartenant à l'ensemble des 190) se trouvent dans ce corpus. Pour chacun de ces mots, nous avons calculé la valeur de la probabilité $p(m / ctx)$, pour lui et ses homonymes.

- Dans 4.958 cas, le mot effectivement dans le texte était plus fréquent que tous ses homonymes. Cela veut dire que dans les autres cas (1116 mots), le système de transcription se serait trompé en choisissant le plus fréquent si l'on avait eu le texte à transcrire.

- Dans 5068 cas, le mot dans le texte était plus probable que tous ses homonymes au sens de la distribution $p(m / ctx)$.

En conclusion, la prédiction par le contexte choisit correctement 110 mots de plus que la simple fréquence. Cela veut dire qu'une faute de ce style sur dix pourrait être évitée. Il faut noter que le travail a été fait sur une portion de tous les mots sémantiquement ambigus du dictionnaire, à savoir seulement 190. Ceci explique le faible nombre (6.074) d'homonymes rencontrés dans les 580.000 mots, ainsi que le pourcentage de fautes (0,2%) auxquelles ils donneraient lieu avec le système de transcription basé sur la fréquence brute. Ce n'est pas en contradiction avec le 1% de fautes de ce genre réellement observé.
Le nombre de mots étudiés et surtout la taille du corpus d'apprentissage, incomplet pour fournir des données contextuelles très fiables, limitent l'étude qui a pu être faite. Il faudrait de plus tenir compte des lemmes plutôt que des formes fléchies. Au niveau d'un corpus limité comme le nôtre, cela n'a pas fait tellement de différence, comme on peut s'en convaincre en regardant les contextes de "mère" ("Annexe I.

Cooccurrences dans le contexte de mère" page 153). Il est rare que plusieurs flexions du même mot apparaissent dans les contextes. Cependant pour des données plus nombreuses, il faudrait compter avec ce phénomène.

Les premiers résultats obtenus montrent que la notion de cooccurrence, traditionnelle en lexicographie, et habituellement utilisée en linguistique descriptive, peut être également féconde en matière de prédiction.

7. CONCLUSION.

Les performances du système ont permis à des sténotypistes de l'utiliser avec profit, couplé à un système de traitement de texte standard pour la correction des fautes et le formattage. Le temps de transcription pourrait être réduit par un facteur 2 au moins par rapport au temps de transcription manuelle.

Après la réalisation de ce prototype de transcription, avec un taux d'erreur de 8.7% et une vitesse de transcription raisonnable, nous avons cherché à améliorer le modèle de langage, au moins sous VM et sans préoccupation de temps machine. C'est l'objet de cette troisième partie.

PARTIE III: MODELISATION SYNTAXIQUE.

La répartition des erreurs avec le modèle markovien, fait apparaître une proportion significative de fautes classées "grammaticales", comme les fautes suivantes (extraites de la transcription donnée in extenso dans "Annexe D. Transcription d'un journal télévisé." page 135)

- *C'est un cardinal... qui a été élu* ***battent*** ...
- *la fumée blanche* ***c' est élevé*** ...
- *si l'élection à la papauté d'un cardinal polonais* ***et*** *une énorme surprise...*
- *l'ascension a été* ***suivi*** ...

Elles sont dues au fait que le modèle, bien que basé sur les classes grammaticales, n'a en mémoire que deux classes et n'opère aucun regroupement des classes précédentes pour garder des constituants comme groupe nominal, verbal etc. Nous avons cherché des outils pour traiter ce genre d'informations. Une première prise en compte de la structure grossière de la phrase (§ III.1 suivante) a donné des résultats encourageants. Ces résultats laissaient espérer qu'il valait la peine d'aller plus loin

dans les contraintes syntaxiques, et d'écrire une grammaire. Après tous les doutes émis sur l'approche grammaticale au §I.3.2, il faudra prévoir des façons d'éviter les inconvénients cités: taille du sous-ensemble du langage naturel accepté, cas de zéro analyse, cas d'analyses multiples. Nos solutions seront exposées dans le §III.3.

1. MODELE STRUCTURAL.

1.1. IDEE INTUITIVE.

Les erreurs grammaticales dans l'expérience du §II.5 comprennent certaines erreurs que nous appelons de "structure". Ceci inclut les fautes résultant dans une phrase ou proposition sans verbe ou sans articulation (conjonction, relatif, etc.), ou trop de sujets pour un seul verbe. Dans les 1800 premiers mots du texte télévisé, nous trouvons 34 erreurs de ce type.
Elles résultent fréquemment d'ambiguïtés entre

à + infinitif / a + participe. (L'artillerie à toucher un réservoir de gaz)

et / est (l'homme et la femme / l'homme est un animal)

Si l'on avait un outil permettant de représenter la phrase comme une suite de constituants de haut niveau, groupe nominal, séquence verbale, groupe prépositionnel, on pourrait alors repérer si la transcription proposée a une articulation correcte, et éliminer par exemple "l'artillerie à toucher un réservoir de gaz" comme n'étant

pas une phrase. C'est ce type de structure que l'heuristique exposée maintenant a pour but d'extraire.

1.2. REPRESENTATION STRUCTURALE D'UNE PHRASE.

Une phrase se décompose en une ou plusieurs propositions principales ou subordonnées. Une proposition simple (déclarative) est en général de la forme:

Sujet Séquence-Verbale (Compléments).

Nous procéderons en deux étapes: découpage de la phrase en propositions simples, puis découpage de chaque proposition en chaînes nominale et verbale.

1.2.1. Découpage d'une phrase en propositions: niveaux.

A chaque phrase est associée une suite de niveaux: Niveau(0), Niveau(1), Niveau(2) ... à valeurs entières 0, 1, 2... On définit ainsi un niveau par proposition apparaissant dans la phrase. Une phrase simple sans subordonnée a un seul niveau 0;

Exemples:

"Il a cité les hommes qui vont parler du sujet qui nous préoccupe"

il a cité les hommes niveau 0

qui vont parler du sujet niveau 1

qui nous préoccupe niveau 2

"La maison que je vois est pleine de gens"

la maison est pleine de gens niveau 0

que je vois niveau 1

La détermination des propositions se fait de gauche à droite au fur et à mesure de la lecture de la phrase du début au point final:

- Le niveau initial est 0.
- Tant que le niveau ne change pas, une proposition se remplit par concaténation des mots rencontrés.
- Le niveau change (ainsi que la proposition courante) lorsque l'on rencontre un mot d'un certain type: il est incrémenté de 1 quand le mot rencontré est un pronom relatif ou une conjonction de subordination. Il est diminué de 1 lorsque le niveau courant est strictement positif et que le mot rencontré est la première frontière (voir ci-dessous) après un groupe verbal.

1.2.2. Découpage d'une proposition en groupes de mots.

Ce découpage est basé non pas sur une description du contenu de ces groupes comme pourrait le faire une grammaire partielle, mais sur le repérage de leurs début et fin (Voir [17]). Certaines classes grammaticales sont dites "frontières" de groupes: les

pronoms sauf les pronoms objets, les verbes, les prépositions, les conjonctions de subordination, les points (début et fin de phrase).

Pour une phrase comprenant n mots m(1...n), supposons que p de ces mots appartiennent à une classe de la liste précédente, ils seront repérés par leur indice dans la phrase: FR_i pour i=1...p.

La phrase peut alors être découpée en groupes de mots G_i pour i=0...p-1, obtenus en concaténant les mots m_j pour $FR_i \leq j < FR_{j+1}$.

En fonction de la frontière gauche et des classes présentes chaque groupe est dit nominal (N), verbal(V) ou prépositionnel (Prep). L'attribution des ces groupes est faite comme suit:

- Quelques règles déterminent si un pronom relatif est sujet, auquel cas il est à la fois une articulation et un début de groupe nominal.
- Dans les autres cas, si la frontière gauche est un point, un relatif non sujet, ou une conjonction de subordination, et qu'un substantif ou adjectif est présent dans le reste du groupe, le groupe est affecté du symbole N.
- Si la frontière est un pronom autre que relatif, le groupe est affecté du symbole N.
- La détermination des autres groupes (V et Prep) est claire.

Exemple:

Phrase:	La maison	que je	vois	est pleine	de gens
Groupes:	N	N	V	V	Prep
Niveaux:	0	1	1	0	0

En procédant de cette façon, la séquence verbale contient non seulement la forme verbale, mais aussi les compléments directs et (ici) les attributs. Ceci n'est pas gênant pour retenir seulement une représentation grossière de la phrase.

1.2.3. Obtention du squelette.

En concaténant pour chaque niveau les symboles N, V rencontrés et en oubliant les symboles Prep, on obtient le "squelette" de chaque niveau.
Exemples.

Pour la phrase ci-dessus, on trouve:

La maison est pleine de gens : squelette(0) = NV

que je vois: squelette(1) = NV

Pour la phrase :

Première réaction ce soir celle de l'archevêque.

Squelette(0) = NN

Pour la phrase :

L'ennemi prit la ville et la pilla.

Squelette(0) = NVV

1.3. INTEGRATION DES DONNEES STRUCTURALES DANS LA TRANSCRIPTION.

1.3.1. Apprentissage.

Bien sûr la structure NV n'est pas la seule que l'on puisse obtenir même pour des propositions correctes. *A priori* n'importe quelle suite de symboles N et V est possible (surtout dans un discours oral). Il est intéressant de regarder leur répartition dans un texte français assez long:

Dans un texte de 20.000 mots (d'origine orale et écrite), en tout 30 structures (squelettes) différentes ont été rencontrées.

NV a été obtenu 817 fois.

V 172 fois.

N 65 fois.

On voit qu'avec des outils simples et sans faire appel à une grammaire, on obtient des informations exploitables, ce qui corrobore les observations faites dans [33]

1.3.2. Nouveau modèle: états(W,S).

Au lieu d'effectuer le choix de la phrase transcrite en maximisant P(W) probabilité de la suite de mots, on peut prédire le couple (phrase,structure) et maximiser P(W,S), qui est prise simplement égale au produit P(W)P(S). P(W) est calculée avec le modèle habituel (partie II), P(S) est estimée par la fréquence relative du squelette dans le corpus d'apprentissage.

1.3.3. Décodage: hypothèses à conserver.

L'estimation de la probabilité de structure ne peut se faire que pour une phrase complète. Comme il a été vu dans la partie II une seule variante est gardée par le système habituel lorsqu'on arrive au point final. Pour éviter une explosion combinatoire, il est en effet exclu de conserver toutes les variantes jusqu'à la fin. Cependant, il est peu coûteux de conserver chaque fois entre deux PPO un petit nombre de variantes avec leurs probabilités par rapport à chacune des précédentes.

Les suites de mots possibles sélectionnées par le préfiltre entre PPO_1 et PPO_2 sont ordonnées par probabilités (modèle de Markov biclasses) décroissantes. Nous conserverons celles qui ont des points frontières différents. Si deux suites ont le même type de frontières, seule celle de meilleure probabilité est à garder. Ainsi, au point final, on a conservé un nombre raisonnable de phrases susceptibles d'avoir des structures différentes. Une énumération permet alors de choisir celle dont le produit P(W)P(S) est maximum.

1.3.4. Résultats.

Ce système modifié a été testé sur les 1800 premiers mots du journal télévisé où le modèle markovien a été évalué.

Sur les 34 erreurs d'articulation, 16 ont disparu.

5 nouvelles erreurs sont apparues, par exemple:

"*Première réaction ce soir sert du représentant...*"

Le bilan est donc largement positif. Les fautes enlevées se répartissent en:

- 5 ambiguïtés *et/est*, par exemple dans la phrase suivante où le mot en gras indique une faute enlevée: "*si l'élection à la papauté d'un cardinal polonais est une énorme surprise...*"
- 2 ambiguïtés *ce/se*, comme dans: "*la foule s' est aussitôt tournée...*"
- 4 ambiguïtés *à + infinitif/a + participe*, comme dans: "*le patronat a proposé la réduction...*"
- 5 autres.

2. GRAMMAIRE PROBABILISTE.

Les informations structurales précédemment décrites donnent de bons résultats, mais le travail pour les raffiner se ramènerait à un certain moment à écrire au moins des parties d'une grammaire (description plus précise des chaînes nominales et verbales). En tout cas, une vraie grammaire devrait faire au moins aussi bien que cette segmentation de phrases. Nous avons donc pris le parti de concevoir une grammaire utilisable pour notre application.

Cependant, une grammaire qui accepte le discours oral sans contrainte, avec un vocabulaire illimité, paraît une tâche non encore résolue. En fait, pour transcrire la sténotypie, ou plus généralement pour la reconnaissance de parole, on peut cerner les caractéristiques souhaitées pour une éventuelle grammaire:

- elle doit accepter une proportion suffisante de phrases, sans pour autant accepter n'importe quoi. Il faut trouver un compromis sur le nombre de règles que l'on va retenir.
- Elle doit essentiellement discriminer les structures de surface "bonnes" ou "mauvaises". Nous n'avons aucune information sémantique disponible dans le dictionnaire, donc nous n'essaierons pas de représenter le sens de la phrase. En particulier il n'est pas nécessaire de rattacher les divers compléments au bon endroit dans la phrase.

- Nous n'essaierons pas non plus de produire absolument une et une seule analyse. La définition de probabilités sur les règles permettra de classer les multiples analyses, ou de fabriquer une pseudo-analyse en raccordant les analyses partielles avec une probabilité faible lorsqu'il n'y a pas de noeud couvrant la totalité de la phrase.

2.1. DESCRIPTION DE LA GRAMMAIRE.

Nous disposions au Centre Scientifique d'un interpréteur de grammaire (GLU, [2]) très commode pour la mise au point. Dans ce système, les règles sont non contextuelles, écrites dans une forme proche de la forme normale de Backus, avec la possibilité de déclarer des attributs attachés à chaque constituant. Ces attributs peuvent être à valeur binaire, entière, ou alphanumérique. Les règles peuvent contenir des conditions sur les attributs, ce qui permet de simuler des règles contextuelles. L'analyseur fournit les analyses trouvées ou les arbres partiels lorsqu'il échoue. Les possibilités d'afficher les traces de l'analyse en font un outil particulièrement puissant pour mettre au point un ensemble de règles. Un extrait de la grammaire est donné dans "Annexe E. Exemples de règles de chaque sorte" page 139.

Les vrais terminaux de notre grammaire sont les 92 classes grammaticales du dictionnaire, plus les mots outils pour lesquels les classes grammaticales n'étaient pas assez précises. Les relatifs, les interrogatifs, certains pronoms, ont été entièrement redéfinis par des règles.
Chaque mot m du dictionnaire est en fait représenté par des règles de la forme p<-m, où p est une classe grammaticale de m.

Environ 200 règles non terminales constituent la grammaire proprement dite. Notons que la possibilité de tester des conditions logiques sur les attributs permet souvent de condenser plusieurs règles en une seule. Nous nous sommes aidée pour les écrire de grammaires traditionnelles [41]. Les constituants de haut niveau sont les non terminaux classiques:

- GN, groupe nominal déterminé, nom propre, pronom personnel ou objet. Le groupe nominal déterminé contient un substantif modifié éventuellement par des adjectifs accordés avec lui, ou simplement un adjectif substantivé, il peut contenir un complément de nom ou une relative. Il peut aussi provenir de la coordination de plusieurs autres. Il possède des attributs genre et nombre. Dans le cas de pronoms, le GN garde l'attribut personne s'il y a lieu.
- VERB, forme verbale composée ou non, comprenant éventuellement la négation, un pronom réfléchi, des clitiques. Un attribut indique l'infinitif.
- VPHRASE, groupe verbal comprenant une forme verbale non infinitive avec des compléments avant ou après et des adverbes. VPHRASE garde les attributs personne, forme composée, passif avec genre et nombre, pronominal, négatif.
- PHRASE, composée en principe de GN et VPHRASE, si les tests d'accord sont satisfaits. Elle peut contenir des subordonnées. Des règles spéciales décrivent les interrogatives. L'interrogation est alors indiquée dans un attribut.
- PSUB, proposition subordonnée composée d'une conjonction et d'une phrase.
- COMPL, complément: préposition et GN, préposition et forme verbale infinitive. Un attribut indique s'il peut être un complément de nom. Pour les mots "au, aux, des, du" le test d'accord est effectué.

De façon générale, la coordination est permise entre constituants de même nature

et produit le même constituant avec les attributs conséquents. Elle a lieu par exemple entre phrases, entre GN, entre VPHRASE (tests de même personne), entre adverbes, adjectifs (tests), subordonnées, relatives, etc.

On se rendra peut-être mieux compte de la puissance de cette grammaire en voyant ce qu'elle n'accepte pas:

- les appositions, par exemple: "Jean-Paul c'est le nom..." Dans un tel cas, la pseudo analyse sera GN PHRASE,
- la coordination entre constituants dissymétriques, par exemple: "nous avons grâce au gouvernement et moyennant votre appui...",
- les impératives, dans ce cas on n'a pas une analyse complète donnant une phrase, mais un seul constituant VPHRASE, cette lacune est due uniquement au fait que les impératifs ne sont pas marqués dans les 92 classes grammaticales,
- l'inversion du sujet dans les phrases comme: "sont reçus les élèves suivants...", Ceci donne également un VPHRASE,
- certains accords avec les verbes comme penser, croire, vouloir etc. dans "elle pouvait être claire",
- certaines ellipses comme: "Quelle cohérence peut-on envisager et quelle formation?"

Dans un certain nombre de ces cas, nous aurions pu rajouter des règles, mais il est apparu que ces nouvelles règles conduisaient à accepter trop de choses et étaient finalement plus nuisibles que bénéfiques pour la transcription. Par exemple, si l'on accepte:

"reine, elle a réalisé..."

il faudra aussi accepter "ouïe Jean-Paul sait..." puisque les virgules ne sont pas marquées dans la sténo.

Nous avons évalué cette grammaire sur un grand ensemble de phrases provenant d'origines diverses (orale et écrite). Environ 45% de ces phrases ont été analysées avec succès (au moins une analyse). Nous considérons ce chiffre comme raisonnable, compte tenu des faits suivants:

- les phrases provenaient en grande partie de discours oral libre, par nature plus complexe et plus souvent "agrammatical" que l'écrit,
- les textes comportaient un certain nombre de fautes de frappe, aboutissant à des faux noms propres répartis un peu n'importe où,
- aucune ponctuation autre que les points n'est utilisée par cette grammaire, puisque la sténo n'en comportera pas,
- enfin certains des textes, sous la forme où nous les avions, n'étaient pas accentués du tout. Les accents ambigus (par exemple chante/chanté) ont été choisis automatiquement par l'étiquetage (§II.2.2). Si les erreurs d'étiquetage (moins de 5%) n'ont pas perturbé grandement les statistiques du modèle de Markov, il est clair qu'au niveau de la phrase les erreurs d'accents peuvent rendre une phrase impossible à analyser.

Une évaluation partielle a montré que parmi les phrases ne comportant pas de fautes de frappe, d'accents, ou de mot inconnu, le taux de phrases analysées est alors 65%. Par ailleurs, malgré ce matériel imparfait, la définition de probabilités sur les arbres partiels permettra de fournir une transcription dans tous les cas, en maximisant la

probabilité de morceaux de phrase, comme nous le verrons dans les paragraphes suivants.

2.2. APPRENTISSAGE DES PROBABILITES.

Pour toute suite de mots $m_1 m_2 ... m_n$, on désire définir la probabilité qu'elle soit produite par la grammaire. Il peut y avoir en réalité plusieurs façons de la produire (analyses multiples). Nous désignons par FULP ("full parse") le noeud initial de la grammaire. C'est le constituant de plus haut niveau. La règle qui le produit exprime qu'un FULP est une PHRASE encadrée par deux points.
Dans le cas où il existe une analyse complète de la suite de mots $m_1 ... m_n$, on notera $r_1 r_2 ... r_n$, la suite de règles de dérivations qui la produit depuis le noeud initial. La probabilité de l'arbre syntaxique associé est le produit des probabilités de chaque r_i.

- Si r_i est une règle terminale, de la forme m<-p, sa probabilité est simplement la fréquence relative du mot m dans la classe p (stockée dans le dictionnaire pour chaque mot).
- Si r_i est une règle non terminale, de la forme A<-BC, nous prenons pour estimé de sa probabilité sa fréquence relative d'utilisation dans un ensemble assez grand de phrases d'apprentissage: chaque phrase est analysée, et s'il y a succès, on met à jour pour chaque règle utilisée le nombre de fois où elle a servi. Quand ces comptes sont obtenus sur le texte entier, on divise le compte de la règle A<-BC

par la somme des comptes de toutes les règles ayant le même constituant de gauche. C'est la probabilité conditionnelle que BC soit dérivé de A.

Remarque:

Jelinek a adapté l'algorithme de Baum (§II.1.2.3) pour ajuster les probabilités des règles d'une grammaire non contextuelle probabiliste. Le fait mathématique sur lequel repose l'assurance de convergence vers la meilleure distribution au regard de l'échantillon observé est encore vérifié. A savoir, la probabilité de l'observation (ici les phrases du corpus) peut encore s'écrire comme une fonction polynomiale homogène des paramètres (probabilités des règles) que l'on cherche à ajuster ([24]). Jelinek l'a appliqué au choix du meilleur arbre quand une phrase est ambiguë pour une grammaire ([24]). Nous avons testé cet algorithme pour ajuster les probabilités des règles, mais les résultats obtenus avec ces probabilités au décodage (expliqué plus bas et dans le paragraphe suivant) sont équivalents ou légèrement inférieurs à ceux obtenus avec les fréquences.

En toute rigueur, la probabilité que la suite de mots soit générée par la grammaire devrait être la somme des probabilités des arbres syntaxiques correspondants lorsque l'analyse est ambiguë. Cependant, pour des raisons de calcul, nous prendrons la probabilité de l'arbre le plus probable, de la même façon que dans le modèle de Markov nous avons remplacé par un maximum la somme intervenant dans l'expression de la probabilité d'une suite de symboles.
Dans le cas où la suite de mots ne peut être couverte par le noeud FULP, l'analyse fournit des arbres partiels disjoints.

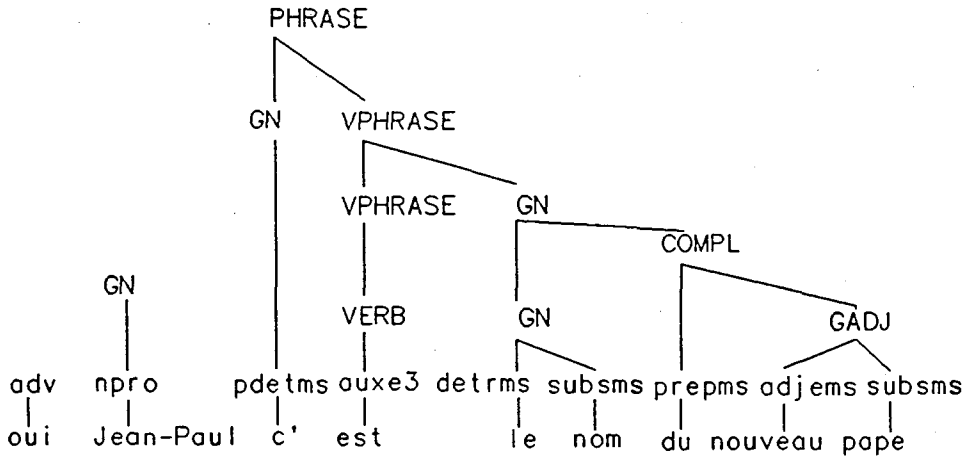


Figure 13: Exemple d'une phrase n'admettant pas d'analyse globale

Chacun des arbres a une probabilité calculée comme précédemment par le produit des probabilités de dérivation. La probabilité de la suite est alors prise égale au produit de celles des arbres et d'un nombre très petit (10^{-4}) pour chaque "trou" entre deux arbres. De cette façon, on pourra quand même maximiser la probabilité de morceaux de phrases même si la phrase entière n'est pas couverte par un FULP. Par exemple, un seul constituant de haut de niveau GN sera automatiquement préféré à une suite de constituants plus bas non reliés entre eux.

2.3. DECODAGE.

Il était difficile d'intégrer l'interpréteur GLU comme une composante du système de transcription. Une fois la grammaire mise au point, seul un analyseur assez rudimentaire est nécessaire. Un tel analyseur fonctionnant sur le principe d'unification, fait maintenant partie du système de transcription (B.Mérialdo). La

grammaire a pu être convertie automatiquement dans le format convenable, tel que:

GN.GE.x.NB.y = DETR.GE.x.NB.y + GADJ.GE.x.NB.y

pour: <GN:1> <- <DETR> <GADJ:GE=GE(1) & NB=NB(1)>

Le décodage se fait phrase par phrase, le principe étant de choisir la transcription de probabilité (au sens de la grammaire cette fois) maximale. Autrement dit, la transcription choisie est celle qui possède l'arbre syntaxique le plus probable.
Comme nous l'avons déjà remarqué, il serait impraticable d'énumérer toutes les transcriptions, en fait les points de passage obligé sont toujours calculés, et seules les phrases composées de groupes de mots sélectionnés par le préfiltre biclasse (partie II) sont envisagées. Au dessus du graphe des mots possibles, sont superposés de façon ascendante tous les noeuds que la grammaire peut produire avec ces mots. Les arbres possibles sont ainsi factorisés au maximum. De plus, si deux noeuds sont construits, couvrant les mêmes mots, avec le même non terminal et les mêmes valeurs d'attributs, celui de plus faible probabilité est éliminé.
Quand tout le graphe est construit, le système énumère tous les arbres et suites d'arbres partiels couvrant toute la sténo, y compris les * du début et de la fin. Il cherche l'arbre ou la suite d'arbres disjoints de plus grande probabilité. Remarquons qu'il préférera un FULP, s'il y en a un, plutôt qu'une suite de non terminaux de plus bas niveaux, parce que le noeud FULP est le seul qui puisse couvrir les points en début et fin de phrase. S'il choisit une suite de noeuds intermédiaires, il doit alors raccorder les deux points en multipliant deux fois par le facteur 10^{-4}, ce qui fait chuter la probabilité.

2.4. RESULTATS.

Le décodage du texte télévisé basé sur ce principe fait disparaître en effet beaucoup d'erreurs classées grammaticales (II.5) que faisait le modèle de Markov. Malheureusement, il fait beaucoup de nouvelles erreurs, soit dans les cas où la grammaire préfère une mauvaise transcription pour avoir une analyse complète (exemple "le matériel est le logiciel"), alors que la phrase oralement prononcée était douteuse, soit surtout dans les cas où plusieurs transcriptions sont grammaticalement correctes, mais la plus probable au sens de la grammaire n'est pas la vraie ("elle l'est probablement au temps en Pologne"). Le taux d'erreurs sur les mots est 10,6%. Le bilan est donc négatif par rapport au modèle markovien, d'autant plus que le temps d'exécution requis par la grammaire est beaucoup plus lourd.

3. COMBINAISON DES APPROCHES MARKOVIENNE ET GRAMMATICALE.

3.1. PRINCIPE.

L'expérience précédente n'est pas très surprenante, puisque les difficultés rencontrées sont exactement celles prévues au début de cet exposé (§I.3.2). En fait le taux d'erreur de 11% est déjà une bonne performance si l'on se rappelle que la grammaire a été écrite à la main alors que les paramètres de la source de Markov ont été ajustés automatiquement. Cette performance montre que la grammaire est assez complète pour l'application.

Les deux approches, markovienne et grammaticale, ont des avantages exactement complémentaires: la source de Markov prédit très bien au niveau du contexte local, tandis que la grammaire réussit au niveau de la structure globale. Il était donc tentant de trouver un moyen de les combiner, qui possède les avantages des deux méthodes. C'est pourquoi nous avons choisi de mélanger les deux probabilités:

- P_t, probabilité au sens des triclasses,
- P_g, probabilité au sens de la grammaire,

d'une suite de mots en faisant leur produit. Ce n'est pas une distribution car la somme n'est pas 1, mais il est équivalent de maximiser ce produit ou de maximiser la distribution qu'on en déduit par normalisation. Ce produit paraît un peu brutal, cela revient à prédire la suite de mots et l'arbre syntaxique de façon indépendante, c'est une hypothèse grossière... Mais intuitivement, nous espérions que l'effet serait de contraindre la grammaire à choisir une analyse complète avec une suite de mots qui soit aussi dans les meilleurs candidats du modèle markovien.

3.2. IMPLEMENTATION.

Pour chaque noeud C construit par la grammaire au dessus des groupes de mots gardés par le préfiltre, on calcule sa probabilité triclasse P_C. Ce nombre est stocké dans la structure le concernant, qui contient aussi les informations sur les constituants dont il provient, sa probabilité grammaticale, et les deux premières et dernières classes sous le noeud. Par exemple pour le groupe nominal "le petit chat gris", la structure associée au GN contient les noeuds dominés DETR et GADJ, et les biclasses (DETRMS ADJEMS), (SUBSMS ADJEMS). Les probabilités triclasses sont calculées de proche en proche, d'abord pour les terminaux (simplement la fréquence mot sur classe). Ensuite pour un noeud C formé à partir de A et B, on calcule P_C comme produit des trois termes:

P_A,

P_B,

probabilité de liaison entre entre A et B, obtenue à partir des deux dernières classes couvertes par A et les deux premières couvertes par B, comme produit des deux triclasses frontières.

Lorsque tout le graphe est construit, on cherche le noeud ou la suite de noeuds qui couvre la totalité de la phrase et qui maximise le produit $P_C \times P_g$. Dans le cas d'une suite de noeuds, il s'agit du produit de ces termes pour chaque noeud, et de ε pour chaque raccordement, comme on l'a vu dans le paragraphe III.2.2.

Il arrive que le graphe soit trop important pour les capacités du programme et que l'on ne puisse conduire l'analyse jusqu'au bout. Cela se produit dans les phrases très longues. Dans ce cas le système donne simplement la transcription triclasse.

3.3. RESULTATS.

Les résultats de ce modèle combiné confirment ces espoirs: le taux d'erreurs, quand le système choisit la transcription de manière à maximiser le produit, tombe à 5,5%. Nous avons comparé les trois performances sur le même texte télévisé. Les noms propres inconnus avaient été mis dans le dictionnaire, la sténotypie est réelle et sans fautes. Les pourcentages d'erreurs avec respectivement le modèle triclasses, la grammaire seule, et le modèle combiné, sont 6,8%, 10,6%, 5,5%. Voici un exemple qui montre bien la différence entre les trois modèles:

- avec le modèle de Markov, 3 erreurs (en gras) sont faites:

  "*La foule* ***c'*** *est aussitôt* ***tourné*** *vers le balcon des appartements du pape où à une heure l'archevêque désormais Jean-Paul est* ***apparue*** *pour le salut traditionnel*"

- avec la grammaire, 4 erreurs sont faites:

  "*La foule s'est aussitôt tournée* ***faire*** *le balcon des appartements du pape* ***ou*** *à une heure l'archevêque désormais Jean-Paul* ***et apparut*** *pour le salut traditionnel*".

- avec la combinaison , une seule erreur est faite:

  "*La foule s'est aussitôt tournée vers le balcon des appartements du pape* ***ou*** *à une heure l'archevêque désormais Jean-Paul est apparu pour le salut traditionnel*".

L'annexe D montre la transcription du journal télévisé avec les différences de transcription entre modèle markovien seul et la combinaison.

CONCLUSION.

Nous avions commencé cet exposé en rappelant les différents niveaux linguistiques concernés par la reconnaissance de parole: lexical, syntaxique, sémantique. La modélisation markovienne peut intervenir aux deux étages syntaxique et sémantique, mais localement. La source que nous avons élaborée pour la sténotypie opère au niveau syntaxique, tandis qu'un modèle trigramme comme celui de Jelinek utilise également de façon cachée des propriétés sémantiques, toujours au niveau local. Cependant, alors que l'on ne dispose d'aucune méthode pour traiter la sémantique d'un domaine ouvert, il en existe au niveau syntaxique, à savoir les grammaires. Il était difficile de dire *a priori* si, de l'approche probabiliste locale ou de la grammaire, l'une se révélerait très supérieure. La différence de performances observée est due à notre avis à la très grande précision obtenue automatiquement pour le contexte local. Pour la même précision en grammaire, il faudrait rajouter beaucoup de règles qui, elles, ne s'écrivent pas automatiquement...

Plutôt que de nous arrêter à la comparaison de deux méthodes, nous avons voulu profiter de cette chance d'avoir deux techniques tout à fait complémentaires en les faisant travailler ensemble. Le taux d'erreur de 5,5% obtenu avec le modèle final

combiné montre que cette démarche était la plus profitable. La performance est réellement très bonne, si l'on se rappelle l'absence totale de contraintes posées sur le dictionnaire et sur le type de discours. Elle nous apparaît comme ce qui peut être obtenu de mieux pour la tâche fixée, si l'on fait intervenir les différents niveaux linguistiques jusqu'au syntaxique inclus.

Le travail pour "gagner" quelques pour cent sur le taux d'erreurs du modèle Markovien était plus important bien sûr que celui fourni pour arriver à ce taux, mais il se trouve largement justifié, d'une part par la performance elle-même, et d'autre part parce qu'il a permis de rapprocher deux méthodes traditionnellement opposées: au lieu de trancher en faveur de l'une d'elles, nous avons tenté de "réconcilier" l'approche probabiliste prédictive, si loin de l'intuition linguistique, et l'approche grammaticale, soucieuse de cerner la réalité des faits linguistiques.

ANNEXE A. LISTE DES CLASSES GRAMMATICALES

Pour la présentation de la liste, on a indiqué par:

la lettre x la marque de genre: quand elle est présente, elle prend les valeurs "M" pour masculin, "F" pour féminin (sauf pour les PPER, pronoms personnels, qui n'ont la marque de genre qu'aux troisièmes personnes),
la lettre y la marque de nombre: elle prendre les valeurs "S" pour singulier, "P" pour pluriel,
la lettre z la marque de personne: elle prend les valeurs 1, 2, 3 pour les première, deuxième, troisième personne du singulier, et 4, 5, 6 pour les première, deuxième, troisième personnes du pluriel.

AAAA	fin de phrase
ADJExy	adjectif
ADJIxy	adjectif indéfini
ADVE	adverbe

AUXA	auxiliaire avoir infinitif ou participe présent
AUXAz	auxiliaire avoir conjugué
AUXE	auxiliaire être infinitif ou participe présent
AUXEz	auxiliaire être conjugué
CCOO	conjonction de coordination
CHIF	nombre cardinal
CSUB	conjonction de subordination
DETRxy	déterminant (défini, indéfini, possessif, démonstratif)
DINTxy	déterminant interrogatif
NE	"ne"
NPRO	nom propre
PAS	"pas"
PAU	"au"
PAUX	"aux"
PDEA	"à" ou "de"
PDES	"des"
PDETxy	pronom démonstratif

PINDxy	pronom indéfini
PINTxy	pronom interrogatif
PPASxy	participe passé
PPERzx	pronom personnel sujet
PPOBxy	pronom objet
PPRE	participe présent
PREFxy	pronom réfléchi
PRELxy	pronom relatif
PREP	préposition
PREPMS	"du"
SUBSxy	substantif
VERBz	verbe conjugué
VINF	verbe infinitif

ANNEXE B. LISTE DES CENT BICLASSES LES PLUS FREQUENTES

DETRFS	SUBSFS	67423
AAAA	AAAA	51156
DETRMS	SUBSMS	47954
PDEA	DETRFS	26390
SUBSMS	PDEA	17926
SUBSFS	PDEA	17835
DETRMP	SUBSMP	16104
SUBSFS	ADJEFS	15722
DETRFP	SUBSFP	12515
SUBSMS	ADJEMS	12052
PREP	DETRFS	11898
CHIF	CHIF	11841
PREP	DETRMS	11452
PDEA	DETRMS	11290
PDEA	VINF	11253
PDEA	SUBSFS	10680
PREPMS	SUBSMS	9962
SUBSFS	AAAA	9301
PDEA	SUBSMS	7825
SUBSMS	AAAA	7814
PDES	SUBSMP	7556

SUBSFS	CCOO	7542
DETRMS	ADJEMS	6744
SUBSMP	PDEA	6718
PPER3M	VERB3	6573
SUBSFS	PREP	6565
SUBSFP	ADJEFP	6407
DETRFS	ADJEFS	6354
PDES	SUBSFP	6234
SUBSMS	CCOO	6212
SUBSMS	PREP	6001
ADJEMS	SUBSMS	5901
PDETMS	AUXE3	5759
ADJEFS	SUBSFS	5629
SUBSMP	ADJEMP	5601
CCOO	PDEA	5465
PAU	SUBSMS	5432
PREP	SUBSFS	5375
ADVE	PDEA	5330
AAAA	DETRMS	4885
AAAA	PPER3M	4590
ADVE	ADVE	4545
SUBSFP	PDEA	4521
PREP	DETRMP	4511
VINF	DETRFS	4432
ADJEMS	PDEA	4403
SUBSFS	PDES	4401
AAAA	DETRFS	4245
VERB3	PDEA	4218
VERB3	ADVE	4194
VINF	PDEA	4134

ADVE	PREP	4130
ADJEFS	AAAA	4129
VINF	PREP	4055
ADJEMS	AAAA	4020
PREP	SUBSMS	4005
AAAA	PREP	3999
PPER1	VERB1	3924
CSUB	PPER3M	3906
CSUB	DETRMS	3872
VINF	DETRMS	3737
AAAA	CCOO	3633
SUBSFS	PREPMS	3568
SUBSMP	CCOO	3530
PREP	DETRFP	3515
ADVE	CSUB	3499
AAAA	ADVE	3463
CSUB	DETRFS	3448
ADJEFS	PDEA	3434
PDEA	NPRO	3396
ADJEFS	CCOO	3328
NPRO	NPRO	3316
CCOO	DETRFS	3299
VERB3	VINF	3297
PPASMS	PREP	3294
NE	VERB3	3248
SUBSMS	PDES	3209
VERB3	PREP	3192
ADJEMS	CCOO	3138
CCOO	ADVE	3129
PREFMS	VERB3	3050

SUBSMP	AAAA	3040
CCOO	PREP	3015
PREP	VINF	2988
VERB3	DETRFS	2916
SUBSMS	PREPMS	2913
CCOO	DETRMS	2902
ADJEMS	PREP	2900
AAAA	PDETMS	2888
PDEA	SUBSMP	2841
DETRMP	ADJEMP	2818
AUXE3	ADVE	2803
ADVE	ADJEMS	2781
PPOBMS	VERB3	2736
AUXA3	PPASMS	2722
PPASMS	PDEA	2672
VERB3	DETRMS	2666
CHIF	SUBSMP	2630
SUBSMS	VERB3	2626
AAAA	PPER1	2559
NE	AUXE3	2558
SUBSMP	PREP	2524
PDEA	ADVE	2511
ADVE	DETRFS	2510
ADVE	DETRMS	2509
SUBSFS	ADVE	2507
ADJEMP	SUBSMP	2496
ADJEFS	PREP	2432
SUBSMS	CSUB	2414
SUBSFS	CSUB	2413
PDETMS	PRELMS	2403

SUBSMS	SUBSMS	2385
SUBSMS	ADVE	2374
PPER3M	NE	2370
SUBSFP	AAAA	2329
SUBSMS	DETRMS	2323
PDEA	SUBSFP	2322
VINF	ADVE	2301
PPOBMS	VINF	2255
SUBSFP	CCOO	2252
SUBSFS	VERB3	2191
SUBSFS	DETRFS	2180
AUXE3	DETRMS	2143
SUBSFS	AUXE3	2113
AUXE3	PPASMS	2065
SUBSMS	PRELMS	2043
CSUB	PREP	2029
PPER3M	AUXE3	2008
ADVE	PPASMS	1983
ADJEMP	AAAA	1974
PRELMS	VERB3	1966
ADJEFP	SUBSFP	1965
AUXE3	ADJEMS	1957
ADVE	ADJEFS	1938
SUBSMS	PPASMS	1925
SUBSMS	DETRFS	1911
ADVE	AUXA3	1906
SUBSMS	AUXE3	1892
AUXE3	PAS	1877
AUXE3	DETRFS	1868
ADVE	AAAA	1852

PPER4	VERB4	1851
ADJEMP	CCOO	1844
SUBSFS	PRELFS	1834
PDEA	DETRMP	1834
ADJEMP	PDEA	1819
AUXE3	PDEA	1817
VINF	CSUB	1791
PDEA	CHIF	1787
VERB3	CSUB	1777
PPASFS	PREP	1769
VERB3	PAS	1757
SUBSFP	PREP	1732
CSUB	PPER1	1731
ADJEFP	AAAA	1728
PPASMP	PREP	1725
CCOO	CSUB	1713
SUBSFS	DETRMS	1701
VINF	AAAA	1689
AAAA	DETRMP	1686
VINF	DETRMP	1681
ADJEFS	ADJEFS	1671
SUBSMS	NPRO	1648
ADJEMS	CSUB	1598
ADJEFP	CCOO	1592
NPRO	AAAA	1587
PREFMS	VINF	1583
PPER3M	ADVE	1570
PREP	ADVE	1569
ADJIMP	DETRMP	1560
AAAA	PDEA	1560

PAUX	SUBSMP	1546
PREP	CSUB	1538
AAAA	CSUB	1525
PDEA	ADJEMS	1469
PREP	CHIF	1453
PPER1	AUXA1	1446
PPOBMP	VINF	1424
SUBSMP	PPASMP	1417
PDEA	PPOBMS	1412
PPER3M	PPOBMS	1384
CCOO	DETRMP	1383
VERB6	PDEA	1368
VERB3	AUXE	1357
ADJEMS	ADJEMS	1351
AUXE3	PREP	1343
VERB6	ADVE	1323
PDEA	DETRFP	1316
SUBSFS	NE	1312
PPER3M	PREFMS	1312
ADJEFP	PDEA	1309
PREP	ADJEMS	1307
PPER1	NE	1300
SUBSMP	PRELMP	1296
PPER1	PPOBMS	1289
AUXE3	PPASFS	1288
VERB6	PREP	1284
CSUB	DETRMP	1281
SUBSFS	PPASFS	1279
CSUB	PPER4	1271

ANNEXE C. LISTE DES CENT TRICLASSES LES PLUS FREQUENTES

PDEA	DETRFS	SUBSFS	24358
DETRFS	SUBSFS	ADJEFS	12734
DETRFS	SUBSFS	PDEA	12529
PREP	DETRFS	SUBSFS	10557
DETRMS	SUBSMS	PDEA	9894
PREP	DETRMS	SUBSMS	9640
SUBSFS	AAAA	AAAA	9301
PDEA	DETRMS	SUBSMS	9276
SUBSMS	AAAA	AAAA	7814
DETRMS	SUBSMS	ADJEMS	7528
CHIF	CHIF	CHIF	7021
DETRFS	SUBSFS	AAAA	5949
SUBSMS	PDEA	DETRFS	5667
AAAA	AAAA	AAAA	5501
SUBSFS	PDEA	DETRFS	5449
DETRFS	SUBSFS	CCOO	4934
AAAA	AAAA	DETRMS	4885
AAAA	AAAA	PPER3M	4590
AAAA	AAAA	DETRFS	4245
ADJEFS	AAAA	AAAA	4129
DETRFS	SUBSFS	PREP	4042

ADJEMS	AAAA	AAAA	4020
AAAA	AAAA	PREP	3999
VINF	DETRFS	SUBSFS	3972
AAAA	DETRMS	SUBSMS	3940
DETRFS	ADJEFS	SUBSFS	3924
AAAA	DETRFS	SUBSFS	3727
AAAA	AAAA	CCOO	3633
DETRMS	SUBSMS	AAAA	3503
AAAA	AAAA	ADVE	3463
PREP	DETRMP	SUBSMP	3459
DETRFS	SUBSFS	PDES	3441
DETRMS	ADJEMS	SUBSMS	3257
VINF	DETRMS	SUBSMS	3125
CSUB	DETRFS	SUBSFS	3114
PREP	DETRFP	SUBSFP	3093
SUBSMP	AAAA	AAAA	3040
CCOO	DETRFS	SUBSFS	3023
DETRMS	SUBSMS	CCOO	2960
DETRFP	SUBSFP	ADJEFP	2940
DETRMP	SUBSMP	PDEA	2911
AAAA	AAAA	PDETMS	2887
CSUB	DETRMS	SUBSMS	2877
DETRMS	SUBSMS	PREP	2851
DETRFS	SUBSFS	PREPMS	2819
SUBSFS	PREPMS	SUBSMS	2781
SUBSMS	PDEA	SUBSFS	2743
SUBSFS	ADJEFS	AAAA	2626
VERB3	DETRFS	SUBSFS	2621
AAAA	AAAA	PPER1	2559
SUBSFS	PDEA	DETRMS	2547

DETRMP	SUBSMP	ADJEMP	2486
CCOO	DETRMS	SUBSMS	2406
SUBSFS	PDEA	SUBSFS	2346
SUBSFP	AAAA	AAAA	2329
SUBSMS	PREPMS	SUBSMS	2290
SUBSFS	PDES	SUBSMP	2216
ADVE	DETRFS	SUBSFS	2199
VERB3	DETRMS	SUBSMS	2190
DETRFP	SUBSFP	PDEA	2154
SUBSMS	PDEA	DETRMS	2078
ADVE	DETRMS	SUBSMS	2075
SUBSFS	ADJEFS	CCOO	2039
SUBSFS	ADJEFS	PDEA	2005
ADJEMP	AAAA	AAAA	1974
SUBSMS	ADJEMS	AAAA	1973
AAAA	PDETMS	AUXE3	1939
PDEA	VINF	DETRFS	1908
SUBSFS	DETRFS	SUBSFS	1873
ADVE	AAAA	AAAA	1852
PDES	SUBSFP	ADJEFP	1840
SUBSMP	PDEA	DETRFS	1798
DETRMS	SUBSMS	PDES	1789
SUBSMS	PDEA	SUBSMS	1774
PREPMS	SUBSMS	ADJEMS	1758
NE	AUXE3	PAS	1746
PDEA	DETRFS	ADJEFS	1745
SUBSMS	DETRFS	SUBSFS	1731
ADJEFP	AAAA	AAAA	1728
SUBSMS	DETRMS	SUBSMS	1727
DETRFS	SUBSFS	ADVE	1722

DETRMS	SUBSMS	VERB3	1717
DETRFS	SUBSFS	VERB3	1692
VINF	AAAA	AAAA	1689
AAAA	AAAA	DETRMP	1686
AAAA	PPER3M	VERB3	1683
PAU	SUBSMS	PDEA	1682
DETRFS	SUBSFS	AUXE3	1676
SUBSFS	CCOO	PDEA	1664
AUXE3	DETRMS	SUBSMS	1661
DETRMS	SUBSMS	PREPMS	1611
SUBSFS	PDES	SUBSFP	1606
SUBSMS	ADJEMS	PDEA	1595
NPRO	AAAA	AAAA	1586
PDEA	DETRMP	SUBSMP	1583
AAAA	AAAA	PDEA	1560
AUXE3	DETRFS	SUBSFS	1544
AAAA	AAAA	CSUB	1525
SUBSMS	ADJEMS	CCOO	1519
PDEA	VINF	DETRMS	1495
PREPMS	SUBSMS	PDEA	1494
PDEA	SUBSFS	AAAA	1489
VINF	DETRMP	SUBSMP	1488
CSUB	PPER3M	VERB3	1480
DETRFS	SUBSFS	DETRFS	1468
SUBSFS	PDEA	VINF	1439
CCOO	PDEA	DETRFS	1424
PDEA	DETRMS	ADJEMS	1423
PDES	SUBSMP	ADJEMP	1419
SUBSFS	PDEA	SUBSMS	1418
PDEA	SUBSFS	CCOO	1391

AAAA	DETRMP	SUBSMP	1374
SUBSMS	PDES	SUBSMP	1371
DETRMP	SUBSMP	CCOO	1367
DETRFS	SUBSFS	PRELFS	1360
SUBSFS	DETRMS	SUBSMS	1359
SUBSFS	PREP	DETRFS	1335
DETRFS	SUBSFS	CSUB	1335
ADJEMS	PDEA	DETRFS	1335
PPER3M	ADVE	AUXA3	1333
PREPMS	SUBSMS	AAAA	1331
SUBSMS	PDEA	VINF	1328
SUBSFS	ADJEFS	PREP	1328
PDEA	SUBSFS	PDEA	1325
DETRMS	SUBSMS	AUXE3	1319
PREP	DETRMS	ADJEMS	1311
NE	VERB3	PAS	1300
PREP	SUBSFS	PDEA	1290
DETRMS	SUBSMS	PRELMS	1289
SUBSMS	PDES	SUBSFP	1284
DETRMS	SUBSMS	ADVE	1280
ADJIMP	DETRMP	SUBSMP	1274
ADJEMS	SUBSMS	PDEA	1273
PDEA	SUBSFS	ADJEFS	1263
SUBSMP	PDEA	SUBSFS	1261
PDEA	VINF	PREP	1253
DETRMS	SUBSMS	SUBSMS	1250
PDEA	VINF	PDEA	1222
PDES	SUBSMP	PDEA	1220
DETRMS	SUBSMS	PPASMS	1211
SUBSFS	PREP	DETRMS	1205

PREP	DETRFS	ADJEFS	1200
PDEA	DETRFP	SUBSFP	1190
SUBSMS	PREP	DETRFS	1187
ADVE	PDEA	DETRFS	1187
ADJEFS	PDEA	DETRFS	1183
AAAA	PPER1	VERB1	1174
VINF	PDEA	DETRFS	1168
SUBSFS	CCOO	DETRFS	1138
PPER3M	VERB3	VINF	1134
PPASMS	DETRFS	SUBSFS	1118
PPER3M	PREFMS	VERB3	1113
PDEA	CHIF	CHIF	1113
CCOO	DETRMP	SUBSMP	1112
VINF	DETRFP	SUBSFP	1110
SUBSMS	PREP	DETRMS	1110
DETRFS	SUBSFS	DETRMS	1085
DETRMP	SUBSMP	AAAA	1072
DETRMS	SUBSMS	CSUB	1070
VERB3	PDEA	VINF	1065
SUBSFP	PDEA	DETRFS	1059
ADJEFS	CCOO	ADJEFS	1049
SUBSFP	ADJEFP	CCOO	1025
DETRFS	SUBSFS	NE	1024
SUBSMS	ADJEMS	PREP	1022
DETRFS	SUBSFS	PPASFS	1020
SUBSFP	ADJEFP	AAAA	1019
CHIF	CHIF	SUBSMP	1019
SUBSMS	CCOO	PDEA	1018
PPER3M	VERB3	ADVE	1009
ADJEFS	SUBSFS	PDEA	1008

DETRMP	SUBSMP	PREP	992
AAAA	AAAA	PPER4	987
PPASMS	AAAA	AAAA	986
SUBSMP	PREPMS	SUBSMS	980
DETRMS	SUBSMS	DETRMS	979
AAAA	AAAA	DETRFP	972
VERB3	PDEA	DETRFS	964
PDEA	SUBSMS	PDEA	964
PPASMS	DETRMS	SUBSMS	957
CSUB	DETRMP	SUBSMP	954
SUBSFP	PDEA	SUBSFS	927
PDES	SUBSFP	PDEA	920
DETRMP	ADJEMP	SUBSMP	918
PPER3M	NE	VERB3	915
PREPMS	SUBSMS	CCOO	914
SUBSMS	CHIF	CHIF	911
DETRFP	SUBSFP	CCOO	909
PDEA	SUBSFS	PREP	893
AAAA	DETRFP	SUBSFP	891
SUBSMP	ADJEMP	AAAA	889
SUBSMP	ADJEMP	CCOO	886
PPER1	VERB1	CSUB	869
PDEA	SUBSMS	AAAA	862
VERB3	AAAA	AAAA	861
PDETMS	AUXE3	PDEA	859
PDETMS	AUXE3	DETRFS	854
SUBSMS	CCOO	DETRMS	845
DETRFP	SUBSFP	AAAA	831
ADJEMS	DETRMS	SUBSMS	830

ANNEXE D. TRANSCRIPTION D'UN JOURNAL TELEVISE.

Le texte qui suit provient de plusieurs journaux télévisés présentés par Roger Giquel (1800 mots). La sténotypie était réelle et sans faute, les noms propres mis dans un dictionnaire utilisateur. Le texte est celui transcrit par le modèle markovien seul. Chaque fois que la transcription par le modèle combiné triclasses et grammaire est différente, elle est donnée en italiques entre parenthèses.

Jean-Paul deux succède à Jean-Paul premier. Oui Jean-Paul c'est (*sait*) le nom du nouveau pape (*bath*) effet exceptionnel dans l'histoire de l'église ce pape n'est pas italien. C'est un cardinal polonais monseigneur Woitliva archevêque de Cracovie qui a été élu battent enfin d'après-midi au deuxième tour de scrutin. C'est une énorme surprise. Le dernier pape non italien dans l'histoire de l'église fût un hollandais Adrien élu en mille neuf cent. Autres éléments de cette énorme surprise est bien sûr que le nouveau pape et (*est*) un archevêque des pays de l'est la Pologne. C'est faire (*vers*) une heure que la fumée blanche c'est élevé (*s'est élevée*) au-dessus de la chapelle sixtine après trois jours de conclave. La foule c'est aussitôt tourné (*s'est aussitôt tournée*) vers le balcon des appartements du pape ou à une heure l'archevêque désormais Jean-Paul est apparue (*apparu*) pour le salut traditionnel. Voici sa première bénédiction prononcée de la logis appelasse (*logis à place*) Saint-Pierre il y a environ vingt minutes. Ce nouveau pape Jean-Paul est âgé de cinquante huit ans. Pour mieux le connaître nous rejoignons à Rome Claude Brovelli qui comme tous les spécialistes a vécu ce moment historique. Le cardinal ne figurait dans aucun pronostic. Déjà il nous faut faire une analyse sur la signification de l'élection de Woitliva. Le fait qu'il est pris le nom de Jean-Paul deux signifie vraisemblablement qu'il inscrit son action dans la continuité de celle de Jean-Paul premier et de bord six. En outre il a joué un rôle appréciable au concile Vatican deux. Qu'en pense ton arôme Bernard chevalier. Si l'élection à la papauté d'un cardinal polonais et (*est*) une énorme surprise pour nous elle les (*elle l'est*) probablement autant en Pologne et dans les pays de l'est. Pour connaître les premières réactions à Cracovie nous avons appelé le

correspondant de l'AFP dans ce pays. Première réaction ce soir celle du représentant des évêques de France monseigneur Etchegarray archevêque de Marseille. Bien entendu notre édition de minuit enfin de programme sera largement consacrée à l'élection de Jean-Paul. Daniel Duigou aura pu rassembler davantage de documents sur le nouveau pape.

Depuis cinq ans depuis qu'Hylary et Tensing ont vaincu le plus au sommet du monde à huit mille huit cent quarante maîtres une vingtaine d'expédition ont suivi leur trace au total cet alpiniste ton deux femmes. Mais les français n'y avait (*avaient*) pas encore planté leur drapeau. En soixante quatorze l'expédition dont faisait parti Gérard Devoissou avait été prise dans une avalanche. Le guide de Chamonix et cinq sherpas furent victime de cette tragédie. D'autre expédition française se sont attaqués à des sommets de l'Himalaya a plus de huit mille maîtres. N'oublions pas que le premier a été vaincu par Maurice Herzog. Mais je le répète il manquait le drapeau français au sommet de l'Everest. C'est fait depuis dimanche matin. Mazeau Afanassief et Geiger ont réussi l'ascension longtemps compromise par le mauvais temps. Ils ont été précédés de peu par trois alpinistes allemands avec lesquels il avait souvent fait cause commune dans la première partie de l'ascension. L'originalité de l'expédition française tient en deux choses l'âge de son chaise il a quarante huit ans et je ne vois pas beaucoup de gens de sa génération être capable d'aller sur le doigt du monde. La deuxième originalité c'est (*sait*) que l'ascension a été suivi (*suivie*) heure par heure par la radio et la télévision en l'occurrence France inter et TF1. Pierre Mazeaud a donc pu dire à la télévision sa joie d'avoir vaincu l'Everest à son tour. Je vous rappelle que pour TF1 c'était Brincourt qui participait à l'expédition avec une équipe complète de tournage. C'est lui que vous allez entendre vous pourrez voir enfin d'années le film complet de cette victoire grave à la caméra de gens Odin qui faisait lui aussi parti de l'expédition.

Un questionnaire bas comme les autres vous sera présenté tout de suite après ce journal. Vous connaissez sans doute les missions de Servan Schreiber présentait (*présenté*) tous les lundis soirs. Ces interviews s'en ignoraient (*sans ignorer*) l'actualité s'attache (*s'attachent*) à aller au fond des choses. L'invité de questionnaire sera ce soir le président de la république. Ce sera donc l'événement du jour à la télévision et le questionnaire sera retransmis en direct de l' Elysée. M. Giscard d'Estaing parlera des chances de la France et de son rôle dans un monde en mouvement.

Même si le climat social est plutôt sombre par les temps qui courent certaines négociations ou propositions de négociations restent à l'ordre du jour. Aujourd'hui par exemple patronat et syndicats négociés (*patronna et s'indiqua négocier*) sur la réforme du système d'indemnisation du chômage. Le patronat à proposer (*a proposé*) la réduction par palier trimestriel des quatre vingt pour cent du salaire en cas de licenciement économique. Cela pourrait aller jusqu'à soixante pour cent au dernier trimestre. En contrepartie le CNPF propose un relèvement des allocations de chômage les plus basses. Et puis demain le patronat et les syndicats font se retrouver pour discuter de l'aménagement de la durée annuelle du travail . Il ne s'agit pas

au moins dans un premier temps de diminuer la durée totale du travail. C'est ce que regrette (*regrettent*) d'ailleurs les syndicats. Il s'agit de mettre sur pied un nouveau système de répartition entre le temps du travail et le temps des loisirs. Des responsables patronaux ont déjà étudié différentes formules en se passant sur le fait quand moyenne un salarié bénéficie de cent trente jours de repos par en.

Ce sont les milices chrétiennes qui ont déclenché la bagarre oliban. Voilà ce cas (*chat*) déclaré notre ministre des affaires étrangères à (*a*) l'issue d'un déjeuner de la presse anglo-américaine. Il faut voir où sont les responsabilités à dire dit. Et elles sont du côté des milices de M. Chamoun. Ce sont elles qui ont déclenché les hostilités elle s'était préparée à ce combat . M. de Guiringaud a tout de même ajouté. Je ne veux pas exonérer les syriens il est vrai qu'ils ont ré agit (*réagi*) très durement. En effet vous vous souvenez de l'épouvantable densité des bombardements de Beyrouth Est jusqu'à ce que l'adresse (*la trêve*) intervienne. Aujourd'hui encore la population de ses quartiers malgré l'adresse vidant (*malgré la trêve fit dans*) les pires conditions de subsistance et de santé. Aujourd'hui encore l'artillerie syrienne à toucher (*a touché*) un réservoir de cave. Et on entendait dédire de franc-tireur (*dès dire de franc-tireurs*) à tous les coins de rue. Pendant ce temps là à vingt kilomètres de Beyrouth il y avait la conférence des pays arabes qui participent physiquement ou financièrement à la force arabe de dissuasion jusqu'ici presque exclusivement syriennes. Non le président algérien où a ri Boumedienne n'a pas disparu. Depuis quelques jours les observateurs s'inquiétaient. On ne l'avait pas vu officiellement depuis près de trois semaines. Le mystère vient d'être levée (*levé*). M. Boumedienne et (*est*) à Moscou. La télévision soviétique a fait parvenir (*avait par venir*) au monde entier les images de sa rencontre avec Leonid Brejnev et Alexis Kossyguine. On se demande naturellement si sa annonce un rat proche ment (*ça annonce un rapprochement*) entre l'URSS et l'Algérie.

Jean-Paul deux a prononcé ce matin son premier message au monde entier à la fin de la messe par laquelle s'achevait le conclave à la chapelle sixtine. Ce message au monde prononcé devant les cardinaux qui la veille l'avait élu pape constitue en quelque sorte une première définition de sa ligne politique sinon un programme de action. Et l'on voit déjà qu'à Jean-Paul premier le pape du sourire qui illumina d'un instant trop bref l'église universelle succède un Jean-Paul deux qui avant même son couronnement dimanche prochain annonce même si c'est en souriant lui aussi qu'il sera le pape de la fermeté sur les principes. Il dit il faut il dit nous devons. Il demande l'obéissance. Il faut appliquer les directives du concile Vatican deux. Sauvegarder l'unité de le église. Reprendre en main la grande charte du concile. Il faut faire face au danger qui menace certaines vérités de la fois catholique. Il ne lui semble pas possible que le drame de la division des chrétiens continus (*continue*). C'est un objet de perplexité et de scandale. Nous allons voir tous ce là avec nos envoyés spéciaux à Rome Claude Brovelli et Bernard chevalier. Mais notez bien que déjà aujourd'hui au Vatican l'événement réellement historique qui avait qu'un cardinal polonais est devenu pape est passé provisoirement au second blanc. Mais sur cela aussi nous allons revenir car si les pays de l'est la Pologne en particulier bien sûr à observer dans

un premier dans un demi silence embarrassé Edouard Girek ailés dirigeants de la Pologne populaire l'ont rompu pour ce qui les concerne cet après-midi en disant leur profonde satisfaction devant l'importante décision du conclave. Commençons par le message de Jean-Paul deux au monde. Claude Brovelli. Bien entendu d'innombrables messages de félicitations de sympathie d'encouragements sont parvenus au Vatican. Vous me permettrez de privilégier celui du président de la république française. La France s'associent (*s'associe*) à la joie et à l'espérance que fait naître dans le monde chrétien l'ouverture de votre pontificat. Vous savez l'amitié que la France éprouve depuis toujours pour la politique et combien elle a d'admiration et de respect pour l'ardente fois chrétienne de la nation polonaise. C'est donc avec une confiance particulière termine Valérie Giscard d'Estaing que je forme des feux pour le succès pour votre haute mission spirituelle au service de la justice et de Labbé entre les hommes. Et puis je vous ai parlé tout à l'heure du message de félicitations des dirigeants polonais. Ils se disent convaincus que le développement continu (*continue*) des relations entre la Pologne populaire et le saint-siège contribuera en faveur de la paix (*Labbé*) de la coopération et de l'amitié entre tous les peuples. Gérard Saint-Paul a rencontré tout à l'heure l'ambassadeur de Pologne en France. A Varsovie hier soir c'était la fête. La fierté polonaise recevait la plus père et la plus inespérée de ses récompenses et c'était la même chose dans toutes les communautés polonaises du monde à Chicago en Amérique et dans le bassin minier dans le nord en France et ailleurs. Mais il faut peut-être se demander si les polonais de Pologne n'avait (*n'avaient*) pas des raisons particulières d'acclamer le cardinal Woitliva. Il y a un mois en effet une lettre pastorale qui était lu dans toutes les églises et qui dit on était inspiré par le nouveau pape déplorer la censure d'état qui est toujours l'arme des systèmes totalitaires qui paralysent la vie culturelle et religieuse de la nation. Ainsi ajouter cette lettre pastorale des millions de croyant en Pologne sont privés de publication religieuse. De plus il regrettait que trois millions de catéchisme seulement et était édité (*aient été édités*) l'an dernier pour huit millions d'enfant. Cette lettre était d'ailleurs tout-à-fait conforme à une précédente déclaration du cardinal Woitliva.

ANNEXE E. EXEMPLES DE REGLES DE CHAQUE SORTE

Les règles suivantes sont un extrait de la grammaire, dans le format GLU utilisé pour la mettre au point. Les : suivis d'un chiffre n au membre de gauche indiquent que le constituant de gauche hérite de tous les attributs du nième constituant de droite.

REGLES TERMINALES DE DEFINITION DES CLASSES DU DICTIONNAIRE.

```
<ADJ: GE=FEM, NB=PLUR><='ADJEFP';
<ADJ: GE=FEM, NB=SING><='ADJEFS';
<ADJ: GE=MASC, NB=PLUR><='ADJEMP';
<ADJ: GE=MASC, NB=SING><='ADJEMS';
```

DECLARATIONS DES CONSTITUANTS.

```
<PHRASE> CONSTRUCT;
<AAA> CONSTRUCT;        point initial et final
<GADJ> CONSTRUCT;       groupe adjectival
<PROPINF> CONSTRUCT;    groupe avec un infinitif comme noyau
<VPHRAS0> CONSTRUCT;    verb phrase à une personne autre que 1
<VPHRAS1> CONSTRUCT;    verb phrase dont la personne est 1
<VPHRASE> CONSTRUCT;    verb phrase
<COMPL> CONSTRUCT;      complément
<GN> CONSTRUCT;         groupe nominal
<RELQUI> CONSTRUCT;     proposition relative
<PSUB> CONSTRUCT;       subordonnée conjonctive
<MOTINT> CONSTRUCT;     mot ou groupe de mots interrogatif
<ADVC> CONSTRUCT;       adverbe de comparaison plus, moins etc.
```

ATTRIBUTS

```
<ACC> FEATURE LOGICAL;  verbe imposant un accord parce que passif
<GENI> FEATURE LOGICAL; compléments de nom (génitif)
<NEG> FEATURE LOGICAL;  négation ne ... pas
```

<COMP> FEATURE LOGICAL; verbes composés
<PRAL> FEATURE LOGICAL; verbes forme pronominale
<PART> FEATURE LOGICAL; participe passé employé comme adjectif
<INT> FEATURE LOGICAL; marque d'interrogation pour verbe, phrase
<SU> FEATURE LOGICAL; marque présence d'un substantif dans un gn
<CO> FEATURE LOGICAL; marque adj avec degré
de comparaison, ou vphrase
<SUP> FEATURE LOGICAL; plus, moins pour superlatif
<OD> FEATURE LOGICAL; objet direct déjà présent pour les verbes.

REGLES DE LA GRAMMAIRE DU FRANCAIS:

Phrases principales. <FULP><-<AAA><PHRASE><AAA>;

Séquences verbales.

<VPHRAS0:1,+OD><-<VPHRAS0><GN:-PRON & -IND & -OB>;
<VERB:1><-<VERB:-ETAT><GADJ:-PART>;
<VPHRAS0:1,-OD><-<VPHRAS0:+OD><COMPL:-GENI>;
<VPHRAS0:1><-<VPHRAS0:-OD><COMPL>;
<VPHRAS0:2><-<COMPL><VPHRAS0>;
<VPHRAS0:1,-OD><-<VPHRAS0:-ETAT><PROPINF>;
<VPHRAS0:1,-OD><-<VPHRAS0:
+ETAT&NB=SING&PERS=3><PROPINF>;
<VPHRAS0:1,-OD><-<VPHRAS0><ADV>;
<VPHRAS0:2><-<ADV><VPHRAS0>;
Restriction "ne...que"
<VPHRASE:1><-<VPHRASE:+NEG><QUE><GN:-PRON>;
Verbe modifié par "plus...que"
<VPHRASE:1,+CO><-<VPHRASE><ADVC>;
<VPHRASE:1><-<VPHRASE:+CO><QUE><GN:-PRON & -INT>;

Phrase type.

Sujet "je" ou "tu"
<PHRASE:1,INT=INT(2)><-<GN:+INV & -INT><VPHRASE: -
PERS=PERS(1) & NB=NB(1)>;
Sujet 3ème personne ou groupes nominaux et verbes actifs
<PHRASE:1,INT=INT(2)><-<GN:-INV&-CE><VPHRASE&:.-ACC &-
PERS=3& NB=NB(1)>;
Adjonction à gauche
<PHRASE:2><-<COMPL><PHRASE>;

Phrases interrogatives:

Interrogation totale
(*Est ce que l'enfant dort*)
<PHRASE:+INT><-<ESC><PHRASE:-INT>;

(*veux-tu te taire...est il ici...*)
<PHRASE:+INT><-<VPHRASE:+INT>;
Interrogation partielle
<GN:1,+INT><-<DINT><GADJ:GE=GE(1) & NB=NB(1)>;
<MOTINT:1><-<GN:+INT>;
<MOTINT><-<MOTINT><ESC>;
Types de structure interrogatives
<PHRASE:+INT><-<MOTINT><VPHRASE:+INT>;
<PHRASE:+INT><-<MOTINT><GN:-INT & -PRON><VPHRASE:
+INT & PERS=3&NB=NB(2)>;
<PHRASE:+INT><-<MOTINT><VPHRASE: PERS=3& NB=NB(3)><GN:
-INT & -PRON>;
<PHRASE:+INT><-<MOTINT><PROPINF>;

Interrogatives indirectes.

<PSUB:+INT><-<MOTINT><PHRASE>;

Propositions relatives

<GN:1,-CE,-PRON><-<GN><RELQUE:
+ACC & GE=GE(1) & NB=NB(1)>;

Subordonnées.

<PSUB><-<CSUB><PHRASE:-RECUR>;
<PSUB><-<QUE><PHRASE:-RECUR>;
<PHRASE:+RECUR><-<PHRASE><PSUB>;

Groupes nominaux

<GADJ:1,+SU><-<SUBS>;
<GADJ:2,+RECUR><-<ADJ:GE=GE(2) & NB=NB(2)><GADJ>;
<GN:2><-<DETR:GE=GE(2) & NB=NB(2)><GADJ>;

Compléments.

<COMPL:+GENI><-<PDEA><GADJ>;
<COMPL:INT=INT(2)><-<PREP><GN:-PRON>;
<COMPL><-<PREP><PROPINF>;

Prise en compte des adverbes

<ADJ:2><-<ADV><ADJ>;
<ADV><-<ADVC>;
<VERB:1><-<VERB><ADV>;

Groupe à noyau infinitif

<PROPINF><-<VINF><GN:-PRON&-OB>;
<PROPINF><-<VINF><GADJ>;

Formes verbales

Négation des formes verbales
<VERB:2,+NEG><-<NE><VERB:-COMP &-NEG><PAS>;
<VINF:2,+NEG><-<NE><PAS><VINF:-NEG>;
Formes verbales avec pronom préverbal
<VERB:2><-<PPOB:+DIR><VERB:-COMP,-PRAL>;
<VERB:2,+PRAL><-<PREF><VERB:-COMP&-ETAT>;
Formes verbales interrogatives
<VERB:1,+INT,+ACC,GE=GE(2)><-<VERB:PERS=3& -
COMP &-PRAL><PPER&:. -
NB=NB(1) & -INV>;
Règles pour passé composé, plus que parfait et passif des verbes
<VERB:1,+COMP><-<AUXA:-INFAUX &-PARTAUX & -ACC><PPAS>;

Coordination.

<ADJ:1><-<ADJ><CCOO><ADJ:GE=GE(1) & NB=NB(1)>;

Règles lexicales manquant au dictionnaire

<AUXE:+INFAUX:TE><='être;
<AUXE:+PARTAUX:TE><='étant';
<AUXA:+INFAUX:TE><='avoir';
<AUXA:+PARTAUX:TE><='ayant';

ANNEXE F. TRACE DU DECODAGE D'UNE PHRASE

La phrase à transcrire était "N'oublions pas que le premier a été vaincu par Maurice Herzog."

On voit ici, de PPO en PPO, la recherche exhaustive dans le dictionnaire ('***** DICO *****), les groupes de mots sélectionnés par le préfiltre (***** PRE FILTRE*****) avec leur probabilité biclasse, et le groupe de mots choisi finalement avec le modèle triclasse (***** CHOIX *****). On peut remarquer que le préfiltre réduit considérablement le nombre de candidats, sans jamais éliminer le bon mot, et que d'autre part le meilleur groupe de mots pour le préfiltre n'est pas toujours celui retenu.

***** DICO *****

nous nous noues nouent noue noue noue oublions ou où où houx houx houes houe plions

***** PRE-FILTRE *****

1 4.0670E-10(nous PPER4 1.00E-00)(plions VERB4 1.28E-04)
2 6.2630E-11(ne NE 1.00E-00)(oublions VERB4 3.08E-03)
3 6.0846E-13(nous PPOBMP1.71E-01)(plions VERB4 1.28E-04)

***** CHOIX *****

1/. .

***** DICO *****

bâcles bâclent bâcle bâcle bats bats bat bas bas bas bas bas bah bâts bât pas pas pas le le premiers premier

***** PRE-FILTRE *****

1 4.2901E-14(pas PAS 5.63E-01)(que CSUB 5.53E-01)
(le DETRMS 3.82E-01) (premier ADJEMS 8.47E-03)
2 2.1107E-15(pas PAS 5.63E-01)(que CSUB 5.53E-01)
(le PPOBMS 5.54E-01) (premier ADJEMS 8.47E-03)
3 1.9762E-16(bâcles VERB2 3.80E-06)(premiers ADJEMP 7.33E-03)

```
  4 1.6596E-16(bâcles VERB2 3.80E-06)(premier ADJEMS 8.47E-03)
  5 1.3221E-16(pas PAS   5.63E-01)(que CSUB  5.53E-01)
(le DETRMS 3.82E-01)(premiers ADJEMP 7.33E-03)
  6 1.1155E-16(bâcle VERB1 8.77E-06)(premier ADJEMS 8.47E-03)
  7 7.4288E-17(bâcle VERB3 4.13E-06)(premier ADJEMS 8.47E-03)
  8 7.2264E-17(bas ADVE  3.31E-04)(que CSUB  5.53E-01)
(le DETRMS 3.82E-01)(premier ADJEMS 8.47E-03)
  9 4.8872E-17(bâclent VERB6 5.70E-06)(premiers ADJEMP 7.33E-03)

***** CHOIX *****

 2/ne oublions

***** DICO *****

as ah a à aidez aider aidait aidais aidais aidaient aidai aidés
aidées aidée aidé étayent
étaye étaye était étais étais étais étaient
étai étés té été hait hais hais haies haie et
est es eh ay ait ais ais aies aient aie ai "
ait "ais "ais "aies "aie dey des des
des dais dais dès dés dé thés thé tes tes t
ait tais tais taies taie taît tés té
hait hais hais haies haie et est es eh ay a
it ais ais aies aient aie ai "ait "ais "ais "aies "aie

***** PRE-FILTRE *****

  1 4.4897E-11(a AUXA3 5.42E-01)(été AUXE  3.07E-01)
  2 4.3159E-11(à PDEA  3.14E-01)(été AUXE  3.07E-01)
  3 3.8964E-12(à PDEA  3.14E-01)(aider VINF  1.46E-03)
  4 2.5011E-12(à PDEA  3.14E-01)(été SUBSMS 1.35E-03)
  5 3.3899E-13(a AUXA3 5.42E-01)(aidé PPASMS 6.73E-04)
  6 1.7790E-13(a AUXA3 5.42E-01)(était AUXE3 2.40E-01)
  7 1.5175E-13(ah ADVE  6.36E-03)(été AUXE  3.07E-01)
  8 1.1372E-13(à PDEA  3.14E-01)(était AUXE3 2.40E-01)
  9 1.1069E-13(à PDEA  3.14E-01)(étés SUBSFP 2.02E-04)
 10 1.1016E-13(ah ADVE  6.36E-03)(était AUXE3 2.40E-01)

***** CHOIX *****
 1/pas que le premier

***** DICO *****

vaincus vaincues vaincue vaincu vint vins vins vins vingt
vin vains vainct vaincs vaincs vainc vain vînt fins fin
```

fin feints feint feins
feins faims faim culs cul hue hue eut eus eues eue eu eût
"ue "ue bars barres bar
rent barre barre barre bar bals balles ballent balle bal
parts part part pars pa
rs parlent parent pare pare par pals pales pale pal pâles
pâles pâle pâle

***** PRE-FILTRE *****

1 5.5930E-10(vaincus PPASMP 2.80E-03)(par PREP 7.00E-02)
2 4.9042E-10(vaincue PPASFS 1.84E-03)(par PREP 7.00E-02)
3 3.7044E-10(vaincu PPASMS 1.49E-03)(par PREP 7.00E-02)
4 1.0877E-10(vaincues PPASFP 8.98E-04)(par PREP 7.00E-02)
5 1.1211E-12(vaincus PPASMP 2.80E-03)(parlent VERB6 3.23E-03)
6 8.9776E-13(vaincu PPASMS1.49E-03)(part SUBSFS5.32E-03)
7 8.4310E-13(vaincue PPASFS1.84E-03)(part SUBSFS5.32E-03)
8 8.2980E-13(vaincue PPASFS1.84E-03)(pâle ADJEFS2.38E-03)
9 7.5764E-13(vaincus PPASMP2.80E-03)(pâles ADJEMP1.29E-03)

***** CHOIX *****

1/a été

***** DICO *****

Maurice Maux mots mot maux os hauts haut eaux eau aux aulx au "auts
haut rives rivent rive Aire haires haire
hères hère erres errent erre erre erre
elles elle airs aires aire air
ailes aile "aires "aire "ères "ère ères
ère socs socques socque soc

***** PRE-FILTRE *****

1 1.0168E-13(Maurice NPRO 2.02E-04)(HERZOG NPRO 2.02E-02)

***** CHOIX *****

3/vaincu par . .

***** PRE-FILTRE *****

1 7.8501E-12(. AAAA 9.56E-02)(. AAAA 9.56E-02)

***** CHOIX *****

1.1574E-11(Maurice NPRO 2.02E-04)(HERZOG NPRO 2.02E-02)

ANNEXE G. TRANSCRIPTION D'UN DISCOURS SAISI EN SITUATION REELLE.

Neuf heures quinze.

Président *** nous allons commencer avec le quart d'heure habituel de retard.

Nous allons entamer notre forum d'aujourd'hui. Je voudrais d'abord remercier nos adhérents ici présent. De leur fidélité au club international de l'IFG *** puisque nous sont le trente janvier. Il est encore dans de leur pressa des tous nos bons veux.

Vous allez me dire. Ce sont peut-être des voeux pieux mais ils sont très sincère. Croyez le bien pour vous mêmes et vos ans de prise. Nous avons souhaité et je vois que vous avez répondu nombreux car pour l'instant il y a peut-être moins de monde mais il j'en aura beaucoup plus tout à l'heure. Le nom nombre d'inscription et assez important. Nous avons souhaité le premier forum de l'année soit c'en devrait sur l'Europe. C'est un problème à l'ordre du jour. Dont on parle souvent. Soit parce que il y a actuellement des négociations à propos de l'élargissement et nous en par le rond enfin de matinée. Soit parce on sent se dessiner un usage plus courants. Plus important de la monnaie européenne. Ce que long appelle l'ECU *** en général.

Soit pour d'autres raisons et il nous apparut eu dire quand cette fin de janvier quatre vingt cinq. Nous puissions échanger dans ce domaine et surtout vous permettre d'entendre d'une part et puis ensuite de dialoguer avec des personnalités importantes qui. A un titre ou à un autre. Sont impliqués dans cette construction européenne.

ANNEXE H. LISTE DES PAIRES LES PLUS FREQUENTES

Ce tableau donne la liste des paires d'homonymes retenues, classés par ordre de fréquence décroissante.

Pour chaque mot sont données l'orthographe et la fréquence marquée dans le dictionnaire.

fois	35665	voix	20958
mère	17679	mer	7573
voix	20958	foi	6373
coup	19786	goût	5487
vers	4678	fer	4018
pas	22167	bas	4000
blanc	4788	plan	3970
fer	4018	verre	3003
temps	43006	dents	2969
goût	5487	cou	2937
foi	6373	voie	2573
bois	9134	poids	2546
fin	11644	faim	2424
sol	3990	sort	2346
maîtres	2419	mètres	2061
champ	2982	chant	1993
classe	2497	glace	1909
saint	8131	sein	1909
fils	9734	villes	1822
odeur	4047	hauteur	1764
ton	6428	don	1682

doigt	2204	toit	1423
yeux	31731	jeux	1409
garde	3814	carte	1388
bord	4581	port	1385
vrai	15308	frais	1383
point	13634	poing	1332
pente	1290	bande	1278
frais	1383	vrais	1162
corps	18554	col	1149
riche	3056	risque	1095
blancs	2315	plans	1088
coeur	23902	choeur	1069
hôtel	4238	autel	1068
fois	35665	voies	1015
vain	2582	fin	1012
vers	4678	verres	1001
somme	4095	zone	976
doigts	4194	toits	956
fils	9734	vice	939
chefs	2251	chaises	939
pu	13732	bu	933
égard	3062	écart	928
fond	15631	fonds	915
points	2630	poings	878
parties	2177	partis	865
ailes	1724	airs	864
goutte	1216	coupe	840
dessin	1361	dessein	840
tâche	1735	tache	833
amis	6609	amies	831

champs	2827	chants	803
face	7783	vase	776
coups	5389	goûts	776
mots	10536	maux	755
pain	3820	bain	718
parts	930	balles	706
pont	1972	bond	688
cartes	1228	gardes	685
part	13573	balle	685
lieux	1516	lieues	681
mères	692	mers	671
mal	8779	mâle	666
balle	685	barre	666
zèle	697	sel	659
feu	7839	voeu	645
saints	1792	seins	642
russe	1161	ruse	621
clairs	711	claires	610
essence	1837	aisance	605

ANNEXE I. COOCCURRENCES DANS LE CONTEXTE DE MERE

Le tableau suivant montre les mots rencontrés dans le contexte du mot mère, avec pour chacun leur classe, leur nombre total de leurs occurences dans le corpus, le nombre de cooccurrences avec mére, et la répartition de ces cooccurrences dans les quatre fenêtres du contexte. Le mot mère a été rencontré en tout 132 fois.

enfant	subsms	173	81	19 23 19 20
père	subsms	145	65	11 22 17 15
mère	subsfs	132	56	17 11 11 17
famille	subsfs	170	23	3 6 13 1
parents	subsmp	79	21	6 4 6 5
femme	subsfs	227	20	8 4 4 4
garde	subsfs	82	18	3 5 4 6
mineur	adjems	41	18	5 5 4 4
faire	vinf	1082	16	5 3 3 5
empereur	subsms	216	15	4 2 5 4
nom	subsms	210	14	4 4 3 3
peut	verb3	1192	13	2 1 7 3
reconnu	ppasms	41	13	4 4 2 3
vie	subsfs	605	13	4 4 2 3
autorité	subsfs	143	12	3 2 4 3
enfants	subsmp	194	12	3 1 4 4
parentale	adjefs	6	11	1 4 3 3
pourra	verb3	208	11	1 2 4 4
recon-naissance	subsfs	38	11	2 5 2 2
seule	adjefs	188	11	2 4 5 0
temps	subsms	537	11	3 2 3 3

ans	subsmp	596	10	2 2 3 3
droit	subsms	471	10	4 2 2 2
égard	subsms	167	9	3 3 1 2
autre	adjefs	433	9	3 1 3 2
fille	subsfs	62	9	2 4 1 2
mariage	subsms	73	9	1 4 3 1
personne	subsfs	171	9	3 0 3 3
petite	adjefs	83	9	3 3 2 1
reconnaître	vinf	35	9	2 2 2 3
veut	verb3	223	9	0 3 3 3
an	subsms	167	8	3 1 2 2
faut	verb3	752	8	2 2 1 3
grand	adjems	309	8	1 4 3 0
même	adjefs	590	8	1 3 2 2
pension	subsfs	32	8	1 2 4 1
tribunal	subsms	145	8	3 1 3 1
défaut	subsms	47	7	1 2 3 1
devoirs	subsmp	52	7	1 5 1 0
dire	vinf	798	7	1 2 3 1
exercer	vinf	62	7	1 0 3 3
jour	subsms	401	7	1 3 1 2
légitime	adjems	20	7	1 3 1 2
marié	ppasms	17	7	2 3 1 1
naissance	subsfs	47	7	3 3 1 0
naturel	adjems	53	7	2 1 1 3
petit	adjems	121	7	1 3 1 2
épouse	subsfs	23	6	1 4 1 0
état	subsms	1031	6	1 1 2 2
accord	subsms	188	6	0 4 1 1

ANNEXE J. LISTE DES TEXTES D'APPRENTISSAGE

Le corpus de 1.200.000 mots sur lequel les statistiques sont apprises est constitué des textes suivants:

- Discours politiques divers (66.000 mots).
- Articles de journaux (40.000).
- Enregistrements d'émissions de Jacques Chancel (13.000).
- La constitution francaise (300.000).
- Discours de Michel Debré (65.000).
- Discours du Parti Communiste (62.000).
- L'exil (51.600).
- Allocutions du Général De Gaulle (108.000).
- Magazines (31.000).
- Manon Lescaut (15.000).
- Rolla, A. de Musset (8.800).
- Discours de Georges Marchais (12.400).
- Livre "Histoire de l'informatique" de René Moreau (21.000).
- Discours d'Alain Poher (2.400).
- Discours de Georges Pompidou (132.000).
- Discours du Parti Socialiste (95.600).
- Discours du RPR (6.200).
- Discours de Sadate (45.600).
- Discours de Senghor (150.000).
- Journaux télévisés (64.700).
- Discours sur le Portugal (50.000)
- Interview de Valéry Giscard d'Estaing (25.500).

BIBLIOGRAPHIE

[1] "Méthode de sténotypie", Ed. Sténotype Grandjean 15 rue Soufflot 75240 Paris Cedex 05.

[2] "User Language Generator", Manuel SB107352-0. Produits spéciaux IBM.

[3] N. Abramson: "Information theory and coding", McGraw Hill, New-York.

[4] A. Andreewsky, J.P. Binquet, F. Debili, C. Fluhr, Y. Hlal, J.S. Liénard, J. Mariani, B. Pouderoux: "Les dictionnaires en formes complètes et leur utilisation dans la transformation lexicale et syntaxique correcte de chaînes phonétiques", 10èmes journées d'étude sur la parole Grenoble 30 Mai-1 Juin 1979.

[5] R.G. Baker, A.C. Downton, A.F. Newell: "Simultaneous speech transcription and TV captions for the deaf", Processing of visible language 2 NATO Conference series, Plenum Press, New York.

[6] L.R. Bahl, R. Bakis, P.S. Cohen, F. Jelinek, B.L. Lewis, R.L. Mercer: "Recognition of a continuously read natural corpus", ICASSP, Tulsa, Avril 1978, pp 422-424.

[7] L.R. Bahl, R. Bakis, P.S. Cohen, A.G. Cole, F. Jelinek, B.L. Lewis, R.L. Mercer: "Further results on the recognition of a continuously read natural corpus", ICASSP, Denver, Avril 1980, pp 872-875.

[8] L. Bahl, S. Das and al...: "Some Experiments with Large Vocabulary Isolated word Sentence recognition", IEEE International Conference on Acoustics, Speech and Signal Processing, March 1984, San Diego.

[9] L.E. Baum: "An Inequality and Associated Maximisation Technique in Statistical Estimation of Probabilistic Fonctions of Markov Process", Inequalities, Vol 3, 1972, pp 1-8.

[10] I.M. Beard: "How British Palantype reporters are helping the deaf": National Shorthand Reporter, January 1983.

[11] M.H. Block: "Captioning the Oscar", National Shorthand Reporter, August 1982.

[12] M.H. Block, M. Okrand: "Real-time closed-captioned television as an educational tool", American Annals for the Deaf, September 1983.

[13] F. Debili: "Traitements syntaxiques utilisant des matrices de précédence fréquentielles construites automatiquement par apprentissage", Thèse de docteur-ingénieur Paris VII, 1977.

[14] A.M. Derouault, B. Merialdo, J.L. Stehlé: "Automatic transcription of French stenotypy", Linguisticae Investigationes Tome VII, 1983, Fascicule 2, Publisher John Benjamins B. V., Amsterdam.

[15] A.M. Derouault, B. Merialdo, J.L. Stehlé: "Une expérience de transcription automatique sténotypie-Français", Technique et Science Informatiques Vol 2, No 5, Septembre-Octobre 1983.

[16] A.M. Derouault, B. Merialdo: "A dictionary structure for stenotypy to French transcription", Etude du Centre Scientifique IBM-France, No 65, Juillet 1983.

[17] A.M. Derouault, B. Merialdo: "Etude des articulations de la phrase en Français. Utilisation des résultats dans un système de transcription automatique Sténotypie-Français", Etude du Centre Scientifique IBM-France, No 67, Octobre 1983.

[18] A.M. Derouault, B. Merialdo: "Le sous-titrage des émissions de télévision à l'usage des mal-entendants", Premier Colloque Image CESTA, Mai 1984, Biarritz.

[19] A.M. Derouault, B. Merialdo "Recognition complexity with large vocabulary", IEEE International Conference on Acoustics, Speech and Signal Processing, March 1984, San Diego.

[20] A-M Derouault, B Merialdo: "TASF: a stenotypy to French transcription system", 7th International Conference on Pattern Recognition, August 1984, Montreal.

[21] A.M. Derouault, B. Merialdo: "Language modeling at the syntactic level", 7th International Conference on Pattern Recognition, August 1984, Montreal.

[22] A. E. Dugourd: "Descriptif du dictionnaire de formes fléchies du Centre Scientifique", Etude du Centre Scientifique IBM-France (à paraître).

[23] A. Fluhr, C. Morel, F. Neel: "Utilisation d'un système automatique de transcription du code sténotypique en français écrit pour le sous-titrage des émissions de télévision", TELEMAT 83.

[24] T. Fusijaki, B. Greene: "A probabilistic approach for dealing with ambiguous syntactic structure", IBM Thomas J. Watson Reasearch Center, Research report.

[25] M. Gross: "Matériaux linguistiques disponibles au LADL en vue d'applications informatiques", Rapport du Laboratoire d'Automatique Documentaire et Linguistique. (à paraître).

[26] F. Jelinek, A.M. Derouault: "Modèle probabiliste d'un langage en reconnaissance de la parole", Annales des Télécommunications, tome 39, numéros 3-4, Mars-Avril 1984

[27] F. Jelinek, R.L. Mercer, L.R. Bahl: "Continuous Speech Recognition: statistical methods", IEEE Transactions PAMI 5, 1983.

[28] F. Jelinek, R.L. Mercer, L.R. Bahl: "Continuous Speech Recognition: statistical methods", Handbook of statistics, Vol 2, Classification, pattern recognition and reduction of dimensionality, Edited by P.R. Krishnaiah and L.N. Kanal 1982 North-Holland.

[29] F.Jelinek, R.L. Mercer: "Interpolated Estimation of Markov Source Parameters from Sparse Data", Proceedings of the Workshop on Pattern Recognition in Practice, May 21-23, 1980, Amsterdam (Netherland), North Holland Publishing Company.

[30] J.P. Haton, J.F. Mari, J.M. Pierrel, S. Sabbagh: "Représentation et mise en oeuvre de contraintes syntaxiques et sémantiques en reconnaissance du discours continu", Séminaire GALF-AFCET, Rennes, Sept 80.

[31] J.F. Larvoire, B. Merialdo, A.M. Derouault: "Etude des homonymies sémantiques", Etude du centre scientifique IBM France (à paraître).

[32] S.Y. Lu and K.S. Fu: "Stochastic error-correcting syntax analysis for recognition of noisy patterns", IEEE Transactions on Computers, Vol C-26, No 12, December 1977, pp 1268-1276.

[33] B. Maegard, E. Spang-Hanssen: "La segmentation automatique du Français écrit", Documents de linguistique quantitative Edition Jean-Favard.

[34] H. Meloni: "Etude et réalisation d'un système de reconnaissance automatique de la parole continue", Thèse d'Etat. Université d'Aix- Marseille II, 1982.

[35] R. Moreau: "Une méthode de décomposition syllabique automatique", Etudes de linguistique appliquée, Ed. Didier, 1965.

[36] J.W. Newitt, A. Odarchenko: "A structure for real-time stenotype transcription", IBM Systems Journal Vol 9, No 1, 1970.

[37] D. Perrin, J. Berstel, M.P. Schutzenberger: "Théorie des codes", Publication du Laboratoire Informatique Théorique et Programmation, PARIS VII.

[38] J.M. Pierrel: "Etudes et mise en oeuvre de contraintes linguistiques en compréhension automatique du discours continu", Thèse d'Etat, Nancy 1981.

[39] W.H. Tsai and K.S. Fu: "Attributed grammars - a tool for combining syntactical and statistical approaches to pattern recognition", IEEE Transactions on Systems, Man and Cybernetics, Vol SMC-10, No 12, December 1980, pp 873-885.

[40] H. Valabrègue: "La transcription phonétique par ordinateur de mots français écrits", Etudes du développement scientifique IBM-France 1965.

[41] R.L. Wagner et J. Pinchon: "Grammaire du français classique et moderne", Ed Hachette.

[42] W.A. Woods: "Transition Network Grammars for Natural Language analysis", Communications of the ACM, Vol 13, no 10, 1970.

[43] Victor W. Zue, Daniel P. Huttenlocher: "Computer recognition of isolated words from large vocabularies", Trends and Applications 1983.

Compagnie IBM France

Réalisé sur les presses de l'Imprimerie d'Entreprise
USINE DE BOIGNY - ORLEANS – FRANCE

0728-04-1985