DOCUMENT RESUME

ED 370 430 FL 022 208

AUTHOR Marty, Fernand
TITLE Caracteristiques de trois systemes informatiques de transcription phonetique et graphemique (Characteristics of Three Computer-Based Systems of Phonetic and Graphemic Transcription).
INSTITUTION Illinois Univ., Urbana. Language Learning Lab.
REPORT NO TR-LLL-T-19-91
PUB DATE Mar 91
NOTE 23p.
PUB TYPE Reports - Evaluative/Feasibility (142)
LANGUAGE French

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Authoring Aids (Programming); Comparative Analysis; *Computational Linguistics; *French; *Graphemes; Language Rhythm; *Phonetic Transcription; *Pronunciation

ABSTRACT

Three computer-based systems for phonetic/graphemic transcription of language are described, compared, and contrasted. The text is entirely in French, with examples given from the French language. The three approaches to transcription are: (1) text entered in standard typography and exiting in phonetic transcription with markers for rhythmic groupings; (2) text entered in poor typography, exiting as for (1) above; and (3) text entered in poor typography, exiting in standard typography. General considerations in developing such transcription approaches are noted, and details of the transcription method are outlined. These aspects of the three transcription methods are then examined: treatment of words whose pronunciation or spelling is ambiguous; grammatical analysis parameters and methods; contextual analysis; reliability of the linguistic and textual analyses; treatment of foreign words and other linguistic anomalies. (MSE)

Language Learning Laboratory

LLL

TECHNICAL REPORT NO. LLL-T-19-91
MARCH 1991

The School of Humanities

University of Illinois
at Urbana-Champaign

CARACTÉRISTIQUES DE TROIS SYSTÈMES INFORMATIQUES DE TRANSCRIPTION PHONÉTIQUE ET GRAPHÉMIQUE

FERNAND MARTY

LANGUAGE LEARNING LABORATORY
College of Liberal Arts and Sciences
University of Illinois at Urbana-Champaign

Technical Report No. LLL-T-19-91

Caractéristiques de Trois Systèmes Informatiques de Transcription Phonétique et Graphémique

Fernand Marty

Professor Emeritus
Department of French
Computer-based Education Research Laboratory
University of Illinois at Urbana-Champaign

March 1991

Available from: Language Learning Laboratory, University of Illinois at Urbana-Champaign, G-70 Foreign Languages Building, 707 S. Mathews, Urbana, Illinois 61801.

Table des matières

Système de transcription 1
Hétérophones et hétérographes 2

Analyse linguistique 5
Difficultés présentées par l'analyse linguistique 5
Choix de la méthode d'analyse linguistique 6
Description de la méthode d'analyse linguistique 7
Niveaux du système d'analyse linguistique 9
Fiabilité de l'analyse linguistique 10

Analyse contextuelle 13
Description de l'analyse contextuelle 14
Fiabilité de l'analyse contextuelle 14

Conclusion 16
1. Mots étrangers 16
2. Position du signal linguistique 17
3. Ambiguïtés linguistiques 17

Caractéristiques de Trois Systèmes Informatiques de Transcription Phonétique et Graphémique

Le but de cet article est de donner une description générale de l'analyse qui sert de support aux logiciels suivants:

L1 Pour ce logiciel, le texte d'entrée est en typographie standard. Le texte de sortie est une transcription phonétique avec marqueurs pour les groupes rythmiques.

L2 Pour ce logiciel, le texte d'entrée est en typographie pauvre. Le texte de sortie est une transcription phonétique avec marqueurs pour les groupes rythmiques. [Par typographie pauvre, nous entendons une typographie sans accents et sans cédilles, tout en majuscules ou en majuscules et minuscules.]

Pour L1 et L2, les schémas intonatifs sont établis par l'utilisateur en fonction des possibilités du synthétiseur de parole.

L3 Pour ce logiciel, le texte d'entrée est en typographie pauvre. Le texte de sortie est en typographie standard.

Nous désirons considérer ces trois logiciels conjointement parce que les principes d'analyse sont fondamentalement les mêmes et que les différences qu'il faut établir entre ces trois logiciels aident à mieux discerner la nature des problèmes qu'il faut résoudre.

Le linguiste qui désire préparer ce type de logiciel doit:

1. Etablir un système de transcription, c'est-à-dire un système de correspondance entre les formes graphémiques et phonétiques pour les logiciels L1 et L2 et un système de correspondance entre la typographie pauvre et la typographie standard pour le logiciel L3.
2. Etablir une analyse linguistique qui détermine la catégorie grammaticale de chacun des mots du texte d'entrée.
3. Etablir une analyse contextuelle qui, par exemple, permet de déterminer si, en l'absence d'une solution linguistique, le mot *fils* appartient au domaine *famille* ou aux domaines *couture/électricité/radio/*, etc.

Système de transcription

Il s'agit de mettre sur pied un système qui donnera la prononciation ou les prononciations de mots tels que:

L1 appendicite, damner, diagnostic, gemment, antiseptique, fils, notions...

L2 ACCORDEON, AVANCA, BEAUTE, AMENERA, POELE, AVIONS, FILS, PATE, JEUNE...

et l'orthographe ou les orthographes standard des mots en typographie pauvre:

L3 BAPTEME, APERCU, NAIF, FORET, CHASSE ...

Il est possible de préparer un lexique qui contiendrait toutes les formes de tous les mots (y compris les noms propres) d'un dictionnaire donné (le dernier **Petit Larousse Illustré**, par exemple). Ce lexique contiendrait près de 500 000 mots puisque les verbes ont au moins 39 formes différentes (*aimer, aime, aimons, aimerait, aimât, aimant*, etc.) et que les noms et adjectifs peuvent avoir jusqu'à cinq formes (*beau, bel, beaux, belle, belles*). La taille en octets d'un tel lexique dépendrait du système de transcription utilisé (IPA, par exemple), mais il faudrait compter au moins une dizaine de millions d'octets; une telle dimension en rendrait l'emploi difficile pour certains utilisateurs potentiels et poserait des problèmes de transport

d'une machine à une autre. Mais le défaut essentiel d'un système de transcription limité à un lexique serait qu'il ne pourrait pas traiter toutes les formes préfixées puisque cet inventaire est illimité. Le préfixe ANTI, par exemple, peut s'utiliser pratiquement avec n'importe quel mot: ANTIZINC serait compris et prononcé correctement par tout francophone; le mot ANTISIDA, qui est maintenant courant, n'est pas dans le dernier **Petit Robert** (1990) ou le dernier **Petit Larousse Illustré** (1991).

Dans un manuel publié en 1985[1], nous avons décrit un système qui réduit considérablement la taille du lexique (moins de 21000 mots pour L1 et environ 29000 mots pour L2 et L3) et qui permet de traiter les néologismes. Le texte d'entrée suit cette route:

1. Lexique d'expressions
2. Lexique de mots individuels
3. Table des préfixes (et retour à 1 si un préfixe est mis en mémoire)
4. Table des racines
5. Table des terminaisons
6. Mise en mémoire du *s* final et retour à 1
7. Règles de transcription

Ce travail a exigé beaucoup de soin et de temps, mais il n'a pas présenté de difficultés insurmontables. Il a été grandement facilité par le grand nombre de livres et d'articles à ce sujet et par l'existence de dictionnaires inverses et de dictionnaires de rimes.

[Ce système traite correctement les mots et les noms propres d'origine étrangère qui sont familiers à la plupart des personnes de langue française, mais il est évident, par exemple, que nous ne pouvons pas prétendre pouvoir transcrire correctement tous les mots et tous les noms propres qui pourraient se trouver dans un article relatant un voyage en Russie.]

Hétérophones et hétérographes

En établissant ce système de transcription, on rencontre des mots dont la prononciation (L1 et L2) ou l'orthographe (L3) est ambiguë.

Logiciel L1:

A. Sans compter les noms propres et les abréviations, nous n'avons trouvé que 147 mots ambigus; ils se répartissent comme suit:

46 pour le contraste Verbe (première personne du pluriel) - Nom pluriel:

acceptions. adoptions, affections, attentions, collections, concoctions, contentions, contractions, dations, désertions, détections, détractions, dictions, diffractions, éditions, électrocutions, exceptions, excrétions, exécutions, exemptions, infections, injections, inspections, intentions, interceptions, interjections, inventions, mentions, notions, objections, oignons, options, persécutions, portions, prospections, rations, rééditions, réfractions, réinventions, relations, rétractions, sécrétions, tractions, transitions, translations, trillions

26 pour le contraste Verbe (troisième personne du pluriel) - Nom/Adjectif:

affluent, coïncident, confluent, content, convergent, couvent, détergent, divergent, dolent, émergent, équivalent, évident, excellent, expédient, féculent, ferment, influent, insolent, négligent, parent, président, résident, somnolent, talent, urgent, violent

[1] F. Marty, R. Hart, Computer Programs to transcribe French Text into Speech: Problems and suggested solutions, Technical Report No LLL-T-6-85. Language Learning Laboratory, University of Illinois, Urbana, 1985

20 pour le contraste Infinitif - Nom singulier:

bitter, boxer, carter, corner, driver, ester, flipper, interviewer, manager, mater, palmer, piper, placer, porter, poster, reporter, sprinter, squatter, stripper, supporter

55 divers:

as, auto, bêta, bis, bois, bus, cacher, campos, cassis, chut, convient, cossus, donc, est, exprès, fier, fils, flous, forte, gens, haste, hélas, jet, job, lacs, las, lias, lis, lut, minerai, obvient, os, ouïe, pagaye, papas, plus, porto, pressent, pub, punch, ras, raya, rhume, rit, sens, seps, soit, suspense, talus, tous, transit, trias, truste, vis, y

Lorsqu'une majuscule est présente, certains mots peuvent être hétérophones. Nous avons rencontré et placé dans notre lexique les mots suivants: *Ben, But, Condom, Dallas, Damas, Duras, Eu, Forez, Havas, Huez, Job* (3 prononciations), *Lot, Marc, Régis, Riez, Rodez, Singer, Suez*. Pour ces contrastes, l'analyse linguistique ne pose généralement pas de problème:

Eu est une ville de Normandie...
Eu égard à votre requête du ...

Les abréviations présentent un problème particulier car leur prononciation dépend parfois du contexte (*av.* qui peut être *avant* ou *avenue*, *S* qui peut être *sud* ou la lettre *s*). Nous avons dû faire un choix et avons placé dans le lexique d'expressions les combinaisons qui sont les plus fréquentes. Exemples:

av. J. C.	Ch. de Gaulle	lat. N	loc. cit.
B. du R.	km/h	lieut. col.	mat. gr.

et, en plus, nous faisons une analyse contextuelle pour *E, S, N, O*.

B. Cet inventaire appelle les remarques suivantes:

1. Dans certains cas, les deux formes du mot appartiennent à la même catégorie grammaticale (*jet, fils, os, gens, pub, lacs, donc*, etc.), mais la plupart appartiennent à des catégories différentes (*y, est, fier, plus, tous, violent, reporter, bus, as, vis, sens, soit, convient*, etc.).
2. Certaines de ces ambiguïtés existent aussi avec un *s* final (*jets, rhumes*, etc.).
3. Normalement l'ambiguïté disparaît lorsqu'un préfixe est utilisé (*rejet, préparent, revis, démentions*, etc.).
4. Pour presque tous les hétérophones, une prononciation est beaucoup plus fréquente que l'autre (*attentions, urgent, porter, y, vis, soit*, etc.) ou même extrêmement rare (*auto, cacher, cossus, gens, lut, oignons, papas, placer, ras, rit, talent*, etc.).
5. Certaines ambiguïtés supposent une opposition phonologique qui n'est pas faite par tous les francophones (*bêta, bois, minerai*, etc.).
6. Certains mots ne sont presque jamais employés (*campos, haste, obvient, seps*, etc.).

C. Deux hétérophones sont beaucoup plus fréquents que les autres. Dans un texte de 10 000 mots, on trouve environ 160 hétérophones. Sur ces 160:

- environ 110 sont pour le mot *est* et, sauf dans les textes spécialisés, il s'agit presque toujours du verbe.
- environ 40 sont pour le mot *plus* et, dans la plupart des cas, la prononciation est /ply/.
- pour le reste, on trouve un ou deux emplois de mots appartenant aux autres catégories (mais *tous* semble avoir un léger avantage).

Un bon logiciel doit effectuer la disambiguïsation de tous ces mots, mais il est important de constater que, même si nous n'avions aucune règle (si *est* était toujours traité comme verbe, par exemple), le nombre d'erreurs dues exclusivement à une absence totale de règles de désambiguïsation serait généralement inférieur à 5 pour un texte de 10 000 mots.

Logiciel L2:

Le nombre de mots hétérophones passe à près de 5 000; un de ces mots a quatre prononciations (COTE = *cote, coté, côte, côté*).

Le nombre de mots dans chaque catégorie grammaticale varie considérablement:

2 941 dans la catégorie verbe/participe passé (AFFIRME)
1 259 dans la catégorie nom/verbe/participe passé (MERITE)
68 dans la catégorie adjectif/verbe/participe passé (INDIGNE)
131 dans la catégorie adjectif/nom/verbe/participe passé (CHAMPIONNE)

[Ces chiffres ne comprennent pas: (a) les mots qui se composent d'un des préfixes de notre table + un mot de notre lexique (ex.: REAFFIRME), (b) les formes en S final (ex: AFFIRMES, HOMMES, GENTILS), (c) les mots en -ONNE qui sont traités séparément lorsqu'ils n'appartiennent qu'à la catégorie verbe/participe passé (ex.: ABANDONNE).]

et un seul par catégorie pour DE, NE, ENTRE, MAIS, ES, CONTRE, etc.

Il y a des oppositions qui sont particulièrement difficiles à résoudre. Exemples:

ILS SONT INDIGNES.	(indignes/indignés)
C'EST UN BEAU DOUBLE.	(double/doublé)
LE MARCHE EST FERME.	(ferme/fermé)
CHAQUE RETRAITE.	(retraite/retraité)
C'EST TROP SALE.	(sale/salé)
NOUS REVERONS.	(rêverons/révérons)
VOUS DEFEREZ.	(déferez/déférez)

Logiciel L3:

L'inventaire des mots hétérographes est presque le même que celui des mots hétérophones de L2 mais certains hétérophones ne sont pas hétérographes:

FILS, JET, LACS, VIOLENT, AS, VIS, FORTE, etc.

et certains homophones sont hétérographes:

DU (du/dû), OU (ou/où), SUR (sur/sûr), FORET (forêt/foret), FUT (fut/fût), DES (des/dés/dès), MANDAT (mandat/mandât), etc.

Pour résoudre les problèmes posés par les hétérophones et les hétérographes, il faut faire appel à l'analyse linguistique et à l'analyse contextuelle.

Analyse linguistique

Cette analyse linguistique est indispensable puisqu'il faut déterminer la catégorie grammaticale de chaque mot du texte d'entrée afin:

a. de choisir la prononciation correcte des hétérophones:

L1 est, tous, plus, président, notions ...

L2 EST, TOUS, PLUS, PRESIDENT, NOTIONS, AFFIRME, NEGLIGE, MAIS ..

ou l'orthographe correcte des hétérographes:

L3 AFFIRME, NEGLIGE, MAIS, FORET, TACHE, PECHE, GENE ...

b. d'établir pour L1 et L2 les groupes rythmiques et, pour chaque groupe rythmique, de choisir les liaisons qui seront faites et les *e instables* qui seront prononcés.

Difficultés présentées par l'analyse linguistique

L'analyse linguistique — par ordinateur — est particulièrement difficile en français car certains des mots les plus fréquents appartiennent à plusieurs catégories grammaticales. Nous utiliserons les abréviations suivantes:

N	nom	A	adjectif	V	verbe
P	pronom	I	infinitif	C	conjonction
D	déterminatif	Adv	adverbe	Pr	préposition
Pp	participe passé	Ppt	participe présent	Ppt	participe présent
Nm	nom masculin	Ppm	participe passé masculin		
Nf	nom féminin	Ppf	participe passé féminin		

Certains mots appartiennent à deux catégories:

D, P	le, la, les, leur...
Pr, P	en
N, V	Plus de 1000 en typographie standard: place, avions, voile, montre, demande, avance, savons...
N, A	Plusieurs centaines car un grand nombre d'adjectifs peuvent fonctionner comme noms: nouvelle, pauvre, vieux, frais, moyen...
N, I	boucher, avoir, pouvoir, sourire...
N, Pr	sous, vers...
N, Adv	pas, bien, rien ...
N, Pp	été, nécessité...
V, Pr	entre
A, Pr	sauf
P, Pp	tu
etc.	

certains appartiennent à trois catégories:

N, A, V	double, critique, lâche, aveugle, trouble, vide...
N, V, Pp	fait, écrit, conduit, bus..
N, A, Adv	fort, droit, franc..
etc.	

et quelques-uns appartiennent à quatre catégories:

N, A, V, Adv	court, ferme
N, A, Pr, Ppt	suivant
etc.	

En typographie standard, le nombre de catégories ambiguës est d'environ 150. En typographie pauvre, il faut ajouter une centaine de catégories qui n'existent pas en typographie standard:

V, Pr, Pp	ENTRE	il entre; entre nous; il est entré
N, D, C	DES	les dés; des livres; dès demain
Adv, C	OU	où je vais; vous ou moi
Nm, Nf, V, Ppm, Ppf	TRAITE	un traité; une traite; il traite; il a traité; il a traite
Nm, Nm, A, V, Pp	DOUBLE	un double; un doublé; coup double; il double; il a doublé
Nf, A, V, Adv, Pp	FERME	une ferme; idée ferme; il ferme; croire ferme; il a fermé
etc.		

Les catégories qui sont communes aux deux typographies ont généralement plus de mots en typographie pauvre; par exemple, le contraste Verbe/Participe passé a environ six fois plus de mots en typographie pauvre qu'en typographie standard, mais cette différence de quantité ne crée pas un problème supplémentaire puisque les mêmes règles d'analyse peuvent s'appliquer.

Ce qui crée un problème, c'est:

1. le fait qu'il y a davantage de catégories à désambiguïser en typographie pauvre
2. le fait que certains contrastes qui facilitent considérablement l'analyse en typographie standard (*a/à, de/dé, ne/né, des/dés/dès, ou/où, la/là, sur/sûr, mais/maïs, du/dû*, etc.) disparaissent en typographie pauvre. Leur absence fait que le nombre de règles d'analyse qui est d'environ 1 300 en typographie standard passe à plus de 11 500 en typographie pauvre.

Marqueurs de phrase:

Il faut définir les conditions dans lesquelles les caractères:

! ; . ?

sont des marqueurs de phrase. Cette analyse est particulièrement difficile pour le point. Exemples:

MM. Duval et Cros ont lu l'art. 10 dans votre lettre du 15 cour. et ...
Il révoqua l'édit. Louis XIV déclara que ...
Il acheta l'édit. Larousse et ...

Choix de la méthode d'analyse linguistique

L'idéal serait une méthode de type arborescent qui analyserait chaque phrase (chaîne de mots entre deux marqueurs de phrase) dans son entièreté et attribuerait une catégorie grammaticale et une fonction syntaxique à chaque mot.

Nous nous sommes heurtés à deux obstacles:

1. La longueur des phrases: L'analyse arborescente devient quasiment impossible si une phrase a plus d'une centaine de mots. Or, chez certains auteurs (Proust, Beckett, etc.) ou dans certains domaines (sociologie, sciences politiques, etc.), des phrases beaucoup plus longues sont relativement fréquentes.
2. La diversité syntaxique: L'esprit humain opère de telle manière qu'il n'est guère conscient de l'ampleur de cette diversité. En l'absence d'un système qui serait un clone de l'intelligence humaine, nous devons faire appel à un ensemble de schémas et de règles. C'est en tentant d'élaborer cet ensemble que l'on se

rend compte que cette diversité est quasi infinie et qu'il semble toujours possible de construire des paires de phrases dont l'une fera échouer l'analyse:

> Le petit jeune homme qui travaille dans ce vieil atelier construit de pauvres matériaux en peu de temps pour faire face à l'inflation qui dure depuis de longues années doit être renvoyé. (construit = participe passé)
>
> Le petit jeune homme qui travaille dans ce vieil atelier construit de pauvres matériaux en peu de temps pour faire face à l'inflation qui dure depuis de longues années et gagne beaucoup d'argent. (construit = 3e personne présent)

On ne peut guère arguer que ces phrases artificielles ne constituent pas un juste test car des phrases plus complexes sont fréquentes dans les romans et journaux:

> Sur sa lancée, on le voyait même, si d'aventure il survivait à grand-mère, se remarier comme son ami des années d'apprentissage à Paris quand tous deux, vingt ans et sans le sou, assuraient la claque pour assister gratuitement aux concerts, lequel ami, après un rapide veuvage, venait de convoler en secondes et tardives noces avec une annoncée jeunette de tout de même cinquante ans, mais de quoi donner des idées à un grand-père brutalement relevé de son engagement de 1912. (Jean Rouaud, *Les Champs d'honneur*, Prix Goncourt 1990, Les Editions de Minuit, p. 50-51.)
>
> Quant à Tonton, si, plutôt que de leur resservir, comme à ses lycéens triés sur le volet, un «je vous ai compris» déjà usé sous d'autres «chienlits», il leur expliquait tout simplement qu'il est à même de les comprendre, car il connaît leur galère, ils en seraient tout étonnés. (Erik Emptaz, *L'art de traiter le problème par les bandes*, Le Canard Enchaîné, 5 déc. 1990)
>
> Pour certains psychanalystes, le mérite dont il est ici question n'est pas celui d'un talent qui se ferait reconnaître, comme en art, ni même de la constance dans un effort créatif, mais uniquement le mérite du sacrifice, de la «peine» qu'on se donne, de la souffrance, en somme, exigée comme compensation de la puissance, excuse de la faute, rachat du péché avant même qu'il ne soit commis. (Jean Daniel, *Les religions d'un président*, Grasset, 1988, p. 183).
>
> Et cette présentation médiatique, sexiste, elle aussi, dans son genre — la femme marin ne compte que quand elle est une jolie sirène qu'on peut prendre en photo sur le ponton est, elle aussi, un couteau à deux lames. (Sylvie Caster, *Florence Arthaud: La femme voilée*, Le Canard Enchaîné, 21 nov. 1990)

Dans l'incapacité de résoudre tous les problèmes posés par une méthode de type arborescent, nous avons décidé d'utiliser une analyse linéaire, récursive, de gauche à droite et il nous semble que cette méthode, aujourd'hui encore, est probablement la plus fiable et la plus facile à modifier lorsqu'il advient qu'une erreur doit être corrigée.

Description de la méthode d'analyse linguistique

Dans les cas peu fréquents où chaque mot de la chaîne à analyser n'appartient qu'à une seule catégorie grammaticale, le module d'analyse est évidemment contourné:

> Ils travaillent pour mon père.

Pour déterminer la catégorie grammaticale des mots qui peuvent appartenir à plusieurs catégories, il nous faut des informations sûres au sujet des mots qui précèdent ou suivent le mot à analyser.

Ces informations nous sont données:

1. par le lexique:

   Nous avons placé dans le lexique plus de 17 000 mots ou expressions qui n'appartiennent qu'à une seule catégorie grammaticale (*pain, vient, ils, chef-d'oeuvre*, etc.).

2. par la table des terminaisons. Exemples:

   - tous les mots qui se terminent par *-emment* sont des adverbes sauf le verbe *gemment*.
   - tous les mots qui se terminent par *-ez* sont des verbes sauf *chez, nez, lez, fez, merguez*, quelques noms propres, et quelques mots ambigus (*Suez, Forez*, etc.).
   - tous les mots qui se terminent par "consonne + *er*" sont des infinitifs sauf quelques noms propres et environ 200 mots qui sont des noms (*cancer, enfer, ver*, etc.) ou des noms/infinitifs (*boucher, pêcher*, etc.).

   Généralement, cette table traite correctement les néologismes ou créations éphémères (*sandwicherie, chausserie, foultitude*, etc.) puisque les personnes de langue française sont conscientes de la répartition de ces terminaisons et choisiront, pour la création d'un nom, la terminaison *-isme* plutôt que la terminaison *-ez*.

En moyenne, le lexique et la table des terminaisons attribuent une catégorie grammaticale non-ambiguë à environ 60% des mots pour le programme L1 et à environ 45% des mots pour les programmes L2 et L3. Par exemple, un test de 1595 mots, composé de 10 textes d'auteurs différents, a donné les résultats suivants:

	Catégorie grammaticale non-ambiguë	Catégorie grammaticale ambiguë	Non-identifiés par le lexique ou les terminaisons
L1	60,3%	37,3%	2,4%
L2 et L3	45,2%	53,2%	1,6%

L'analyse linguistique concerne donc environ 40% des mots pour L1 et 55% des mots pour L2 et L3.

Les mots qui exigent une analyse linguistique peuvent se trouver isolés (c'est-à-dire entre deux éléments dont la catégorie grammaticale est sûre):

Je connais son mari. (son = nom ou déterminatif)

ou en groupes de 2, 3, 4, etc. mots ambigus consécutifs:

Il était devant la porte de notre maison.

ou dans des phrases où chaque mot appartient à plusieurs catégories grammaticales:

Le président lut le poster.
L'excellent président lut le poster.
L'excellent président y lut le poster.
L'excellent président y lut le violent poster.
L'excellent président y lut exprès le violent poster.
L'excellent président-bis y lut exprès le violent poster.
L'excellent président-bis y lut exprès le violent poster négligent.
etc.

L'analyse doit donc:

- déceler pour chaque mot ambigu isolé toutes les constructions où ce mot peut se rencontrer et déterminer une solution pour chacune de ces constructions; par exemple, quelles sont les constructions où le mot *devant* peut se rencontrer et quelles sont celles où il est une préposition, celles où il est un nom, et celles où il est un participe présent.

- déceler pour chaque chaîne de deux, trois, quatre, cinq, etc. mots ambigus consécutifs toutes les constructions où cette chaîne peut se rencontrer et déterminer une solution pour chacune de ces constructions; par exemple, cette chaîne de quatre mots ambigus:

  déterminatif/pronom + nom/verbe + déterminatif/pronom + nom/verbe

peut parfois être analysée sans faire appel aux mots qui précèdent ou suivent:

la réserve le place

puisque la présence de *le* devant *place* résout le problème. Mais cet ensemble

leur réserve la place

peut être analysé de deux façons suivant le contexte:

Elle leur réserve la place...

Elle dit que leur réserve la place...

Il nous est vite apparu que le nombre de constructions et leur diversité étaient tels qu'il nous serait impossible, *in abstracto*, de percevoir tous les problèmes et que la seule solution était d'établir et de modifier le module d'analyse au fur et à mesure que nous faisions passer des textes par ce module.

Au cours de ces dix dernières années, des milliers de tests ont affiné cette analyse linéaire et l'ont élevée à ce qui semble être le quasi-maximum de ses possibilités; depuis 1987, la majeure partie de ces tests a été effectuée au CNET de Lannion par Madame Danielle Larreur sur des textes d'une grande variété.

Niveaux du système d'analyse linguistique

Après avoir expérimenté plusieurs possibilités, nous avons maintenant un système qui comporte trois niveaux et qui utilise des ensembles de catégories grammaticales. L'ensemble 657, par exemple, contient les mots (verbes, pronoms objet, la négation *ne*, les pronoms relatifs *qui, dont*, etc.) qui, dans un contexte donné, placent nécessairement le mot à analyser dans une catégorie non-ambiguë; cet ensemble, par exemple, permet de résoudre l'ambiguïté pronom/préposition pour le mot *en* dans cette construction:

Le gouvernement en place met/lui/ne/qui...

1. Le niveau A ne s'applique qu'aux homophones et homographes; c'est une suite de règles de ce type:

   356 345 129 378 210 003 456 341 05

   Chaque règle tente de désambiguïser un ensemble de catégories grammaticales (par exemple, tous les mots qui peuvent être nom/verbe/participe passé: *fait, conduit, permis*, etc.). Cette règle examine les quatre ensembles qui précèdent et les trois ensembles qui suivent. La règle se lit donc ainsi:

   Si un mot qui appartient à l'ensemble 210 est précédé de mots qui appartiennent aux ensembles 356, 345, 129, 378 et s'il est suivi de mots qui appartiennent aux ensembles 003 456 341, ce mot est un verbe (05).

   Cette analyse est récursive, c'est-à-dire qu'il est parfois nécessaire d'opérer plusieurs passages pour établir la catégorie grammaticale de tous les mots. Par exemple, la phrase:

   Ce vieux combat offre un espoir.

   exige deux passages. Au premier passage, *ce, combat*, et *offre* sont désambiguïsés; au second passage, il est devenu clair que *vieux* ne peut pas être un nom et il est classé adjectif.

2. Le niveau B est chargé de retoucher l'analyse. Les rares mots qui n'ont pas été désambiguïsés sont placés par défaut dans la catégorie la plus vraisemblable, certains noms communs deviennent noms

propres, certains noms et adjectifs dont le nombre était incertain (*gaz*, *mauvais*, par exemple) sont classés singulier ou pluriel suivant le contexte.

3. Le niveau C s'applique aux hérétophones et hétérographes; les règles déterminent la catégorie grammaticale et, par conséquent, la prononciation ou l'orthographe à utiliser. Puisque l'analyse de certains de ces mots est particulièrement difficile (*convient*) et importante (*est*, *tous*), les règles utilisées à ce niveau examinent les cinq ensembles qui précèdent et les quatre ensembles qui suivent le mot en cours d'analyse. De plus, nous pouvons spécifier certaines caractéristiques de ce mot: présence ou absence de trait d'union, nature de la dernière lettre, majuscule ou minuscule, position du mot dans la phrase, etc.

Fiabilité de l'analyse linguistique:

A. Dans les rares cas où le signal linguistique est trop éloigné, l'analyse linéaire ne choisit pas toujours la forme correcte. Exemples:

   1. Notre programme place *somnolent* dans la catégorie adjectif dans ces deux phrases parce que l'analyse ne perçoit pas que *gens* est le sujet de *somnolent* dans la deuxième phrase:

      Voici les braves gens qui aiment tant les vieux habitants de ce tout petit village somnolent. (Adjectif)

      Mais les braves gens qui aiment tant les vieux habitants de ce tout petit village somnolent. (Verbe)

   2. Notre programme place *menace* dans la catégorie verbe dans ces deux cas parce que le signal linguistique (*ET DIT/DIT*) est trop éloigné.

      LE CHEF DU PERSONNEL MENACE DEPUIS LE DEBUT DES GREVES TOURNANTES EN JUIN DERNIER DE LICENCIER SES EMPLOYES ET DIT QUE .... (MENACE = menace)

      LE CHEF DU PERSONNEL MENACE DEPUIS LE DEBUT DES GREVES TOURNANTES EN JUIN DERNIER DE PERDRE SON EMPLOI DIT QUE .... (MENACE = menacé)

B. Il existe des constructions et des combinaisons de mots qui restent ambiguës parce qu'il n'y a pas de signal linguistique qui permette de résoudre l'ambiguïté. Exemples:

Programme L1:

Les phrases qui acceptent plusieurs analyses linguistiques sont rares, soit parce que le nombre de constructions qui sont toujours ambiguës est très faible:

Devant un tel (Nom), il dit... Devant = préposition ou participe présent

soit parce que l'ambiguïté exige pour se réaliser un ensemble de facteurs dont l'occurrence est statistiquement faible. Par exemple, ces deux syntagmes sont fréquents:

D	+	N	+	V	+	D	+	N
La		fille		regarde		la		télévision.
D	+	A	+	N	+	P	+	V
La		jolie		fille		la		regarde.

mais l'amalgame de ces deux constructions est rare puisqu'il exige une suite de quatre mots appartenant chacun à deux catégories grammaticales:

D	+	A/N	+	N/V	+	D/P	+	N/V

et cet amalgame ne peut être ambigu que s'il y a possibilité d'accord en genre et en nombre pour les trois premiers mots et pour les deux derniers mots:

Le vieux garde // la place.	D	+	A	+	N	+	P	+	V
Le vieux // garde la place.	D	+	N	+	V	+	D	+	N
Cette vieille manie // le charme.	D	+	A	+	N	+	P	+	V
Cette vieille // manie le charme.	D	+	N	+	V	+	D	+	N

L'ambiguïté disparaît lorsqu'il n'y a pas possibilité d'accord:

Le vieux // demande la place.
La vieille // partage le charme.
Le vieux garde // le place.
Le vieux garde // les place.

ou, évidemment, lorsqu'il y a un signal linguistique:

Le vieux // garde la place qui ...

L'ambiguïté linguistique peut aussi se manifester dans les constructions suivantes:

a. D + A/N + A/N/V + D/P + N/V

La belle ferme le voile.	D + A + N // P + V
La belle ferme le voile.	D + N + A // P + V
La belle ferme le voile.	D + N // V + D + N

b. D/P + N/V

Il décrivit le début et le sort de la bataille.	D + N
Il attrape le chien et le sort de la salle.	P + V
Il expliqua le nouveau cours et le but du projet des étudiants.	D + N
Il accepta le nouveau breuvage et le but du bout des lèvres.	P + V
... et celui qui en attribue aux autres le mérite.	D + N
... et celui qui en donne aux pauvres le mérite.	P + V

c. A/N + N/V

Il apporte le philtre et la bonne plante du rebouteux.	A + N // D + N
Il taille la vigne et la bonne plante du persil.	N // V + D + N

d. Pr/P + N/V

Il accuse les prêtres et le ministre en place dans ce pays.	Pr + N
Il recueille les enfants et le ministre en place dans ce pays.	P + V

e. Pr + D/P + N/I

Il est parti sans le boucher.	Pr + D + N
Il est parti sans le boucher.	Pr + P + I

f. A/N + A/N

C'est un pauvre aveugle.	A + N
C'est un pauvre aveugle.	N + A

g. D + N + A/Pp + Pr + D + N

Je passe du texte écrit à la prononciation.	D + N + A // Pr + D + N
Je parle du texte écrit à la réunion. (= qui a été écrit à la réunion)	D + N // Pp + Pr + D + N

Il confie les enfants trouvés à la Croix-Rouge.	D + N + A // Pr + D + N
Il interroge les enfants trouvés à la gare.	D + N // Pp + Pr + D + N
(= qui ont été trouvés à la gare)	

h. D + N + A/Pp + Pr + I

... les marches forcées pour habituer les soldats.	D + N + A // Pr + I
... les moyens déployés pour sauver les baleines.	D + N // Pp + Pr + I
(= qui ont été déployés pour sauver les baleines)	

Et il semble que, si l'on analysait suffisamment de textes, on pourrait trouver une construction ambiguë pour pratiquement chacun des mots qui appartiennent à plusieurs catégories linguistiques:

a. *Entre* (préposition ou verbe):

Il plante les fleurs ici et entre les arbres dans le pré.	(Préposition)
Il copie les chiffres ici et entre les données dans l'ordinateur.	(Verbe)

b. *Sauf* (préposition ou adjectif):

Il est toujours sauf sous l'arbre.	(Adjectif: toujours sauf // sous)
Il est partout sauf sous l'arbre.	(Préposition: partout // sauf)
Il est toujours sauf sous l'arbre.	(Adjectif: toujours sauf // sous)
Il pousse toujours sauf sous l'arbre.	(Préposition: toujours // sauf)

c. *Tout* (adjectif indéfini ou pronom indéfini) + *bien que*, *pendant que*, etc.

Elle est jalouse de tout bien que tu acquiers.	(de tout bien // que tu)
Elle est jalouse de tout bien que tu souffres.	(de tout // bien que tu)

d. Verbe + *trop* + préposition:

Vous travaillez trop // près de la maison.	(*trop* modifie *travaillez*)
Vous travaillez // trop près de la maison.	(*trop* modifie *près*)

e. *Son/Ton* (nom ou déterminatif):

... et la prise de son aide.	(*son* = nom)
... et la valeur de son aide.	(*son* = déterminatif)

f. Formes de *tenir compte*:

... et la seule solution qui tienne compte des difficultés présentes.
... et la seule solution qui tienne // compte des avantages importants.

g. Verbes qui sont singulier et pluriel:

C'est le X des X qui nous convient.	(3e personne singulier ou pluriel)

h. Formes de *trouver grâce*:

La chouette a trouvé grâce à ses yeux.	(a trouvé // grâce à ses yeux)
La pénitente a trouvé grâce à ses yeux.	(a trouvé grâce // à ses yeux)

Programmes L2 et L3:

L'absence d'accents et de cédilles augmente les possibilités d'ambiguïtés linguistiques puisque des oppositions utiles à l'analyse disparaissent: *a/à*, *la/là*, *sur/sûr*, *de/dé*, *du/dû*, *des/dés/dès*, *ou/où*, *ne/né*, *entre/entré*, *arrière/arriéré*, présent/participe passé (*affirme/affirmé*), etc. Exemples:

IL A DU POUVOIR.	(du/dû)
LA FRANCE OU MEME PARIS NE L'ATTIRE PLUS.	(ou/où)
JE M'AMUSE OU JE TRAVAILLE.	(ou/où)
LA VILLE OU IL EST NE LUI DONNE PLUS D'ARGENT.	(ne/né)
IL A PRIS SA RETRAITE ET ACHEVE SA VIE A PARIS.	(achève/achevé)
IL A DEUX DES PIPES.	(des pipes/dés pipés)
QUI EST MARIE?	(Marie/marié)
UN DE NOUS DIRA TOUT.	(de/dé)
CHAQUE LETTRE EN TIENT COMPTE.	(lettre/lettré)
C'EST LE NOM DE CHAQUE FORET.	(forêt/foret)
LE FERMIER A LE GRAIN ET LE VALET A LA PAILLE.	(a la paille)
LE FERMIER A LE MARTEAU ET LE CISEAU A LA MAIN.	(à la main)
LA GARE OU LE TRAIN S'ARRETE.	(où)
L'AUTOBUS OU LE TRAIN S'ARRETE.	(ou)
UN FEU ARRIERE.	(arrière)
UN ENFANT ARRIERE.	(arriéré)
JE SUIS SUR DES CHARBONS ARDENTS.	(sur)
JE SUIS SUR DES RESULTATS PRESENTS.	(sûr)
IL A, COMME TOUJOURS, DONNE SON ACCORD.	(donné)
ET LUI, COMME TOUJOURS, DONNE SON ACCORD.	(donne)
LES FEUILLES, QUI SONT TRES SECHES, DONNENT...	(sèches)
LES BOIS, QUI SONT TRES SECHES, DONNENT...	(séchés)
LE FAIT QUE LE NOMBRE ELEVE DES LAPINS EST...	(élevé)
LE FAIT QUE LE CHEF ELEVE DES LAPINS EST...	(élève)

Nous discuterons dans les deux prochaines sections les solutions qui peuvent être apportées à ces problèmes.

Analyse contextuelle

L'analyse contextuelle peut être utile dans ces trois cas:

1. Elle peut déterminer la prononciation des hétérophones et l'orthographe des hétérographes qui appartiennent à la même catégorie grammaticale:

... mes fils ...	(/fis/ - /fil/)
... LEUR COTE ...	(cote/côté/côte)
... NOUS GENERONS ...	(générons/gênerons)

2. Elle peut aider à résoudre certaines des ambiguïtés linguistiques mentionnées dans la section précédente.

3. Elle peut aider à déterminer le découpage de la phrase en groupes rythmiques dans les cas très rares où l'analyse linguistique est correcte, mais les fonctions restent ambiguës. Exemples:

   Elle avait des fractures // de la tête aux pieds.
   plutôt que: Elle avait des fractures de la tête // aux pieds.

Il faut chasser les clochards // du métro.
plutôt que: Il faut chasser // les clochards du métro.

Il veut protéger les musulmans // de l'Islam.
plutôt que: Il veut protéger // les musulmans de l'Islam.

Nous sommes sortis de ces bas-fonds // estomaqués.
plutôt que: Nous sommes sortis de ces bas-fonds estomaqués.

Heureusement, dès le lendemain, nous avons appris ce qui s'était passé // dans le Figaro.
plutôt que: Heureusement, dès le lendemain, nous avons appris // ce qui s'était passé dans le Figaro.

... un vin blanc // à consommer le soir.
... un objet // propre à satisfaire leur désir.

Description de l'analyse contextuelle

L'analyse que nous utilisons compare le contenu sémantique de la chaîne en cours d'analyse et des deux chaînes précédentes avec des listes de mots généralement associés avec les champs contextuels des mots qu'il faut désambiguïser. Exemples:

... ses fils ...: Si l'analyse trouve deux mots associés avec le sens /fis/ (ex: *aîné*, *marraine*) et un seul avec le sens /fil/ (ex: *laine*), le sens /fis/ est choisi. Le sens /fis/ est aussi choisi si l'analyse trouve le même nombre de mots pour chaque sens.

... LA COTE ...: Si l'analyse trouve plus de mots associés avec le sens *côte* (ex: *falaise*, *azur*, *corniche*) qu'avec le sens *cote* (ex: *cours*, *vente*), le sens *côte* est choisi.

Notre analyse tente aussi de déterminer si le texte est au passé ou au présent ou s'il agit d'un féminin ou d'un masculin:

IL A HESITE, RECULE DE NOUVEAU ET ABANDONNE.	(reculé, abandonné)
IL HESITE, RECULE DE NOUVEAU ET ABANDONNE.	(recule, abandonne)
JEANNE, QUI A ETE INQUIETE, DIT QUE ...	(inquiète)
JEAN, QUI A ETE INQUIETE, DIT QUE ...	(inquiété)

Fiabilité de l'analyse contextuelle

Cette analyse contextuelle est utile, surtout pour les mots où les champs sémantiques sont très différents (*fils, lacs, pub, jet*, etc.), mais elle est rudimentaire par comparaison avec le système employé par l'être humain. Il ne semble guère possible, du moins dans un proche avenir, de programmer tout un ensemble de relations conceptuelles, de déductions, et d'inférences qui pourrait examiner n'importe quelle situation et conclure, par exemple, que la phrase *La belle ferme le voile* décrit une femme qui ferme le rideau parce qu'il fait froid, qu'il fait chaud, qu'il fait du soleil, qu'elle part en vacances, que son voisin la regarde, qu'elle va se déshabiller, qu'elle a des invités, etc., ou que la phrase:

La nouvelle alarme le peuple.

est presque certainement du type:

D + N + V + D + N

car il n'est guère concevable qu'une alarme puisse peupler.

D'autre part, nous n'avons pas pu faire une analyse contextuelle pour toutes les possibilités d'ambiguïtés linguistiques car, s'il est vrai que les phrases qui acceptent plusieurs analyses linguistiques ne se rencontrent que rarement dans les textes, il n'en reste pas moins qu'elles peuvent se présenter et, comme nous l'avons montré, qu'elles peuvent se présenter pratiquement avec n'importe lequel des milliers de mots qui appartiennent aux catégories Adjectif/Nom, Verbe/Nom, Infinitif/Nom, etc. Pour tenter de résoudre toutes les possibilités d'ambiguïté linguistique, il faudrait donc prévoir une analyse contextuelle pour chacun de ces mots dans chacune des constructions où il peut se rencontrer, c'est-à-dire accomplir un travail immense qui exigerait un code de plusieurs millions d'octets.

Un tel travail ne semble pas justifié puisque le programme ne nécessite l'analyse contextuelle que rarement et puisque, dans l'état actuel de nos connaissances, le taux de succès dans certains cas (*boucher*, *INDIGNE*, etc.) serait loin d'être suffisant.

Nous avons donc fait un choix et nous avons limité notre analyse conceptuelle aux cas les plus importants et/ou les plus faciles.

Dans le programme L1, l'analyse contextuelle ne s'applique qu'aux hétérophones pour lesquels l'analyse linguistique est impossible ou risque de rester ambiguë: *bis, campos, cassis, convient, cossus, fils, forte, gens, haste, jet, job, lacs, las, ouïe, pub, punch, suspense, y, Ben, But, Condom, Damas, Eu, Forez, Job, Lot, Marc, N, E, W, S, Rodez, Suez,*

et nous avons placé certaines combinaisons sûres dans notre lexique. Exemples:

Alpe d'Huez	est-il
arrière-petits-fils	il n'est
compte-fils	fier-à-bras
fils de laiton	pied talus
père et fils	vis à vis de

Pour les autres cas, nous tentons de choisir la forme la plus fréquente. Exemples:

le gouvernement en place	(préposition + nom)
un pauvre savant	(adjectif + nom)
sans le boucher	(déterminatif + nom)

Dans le programme L2, en plus des mots cités pour L1, nous traitons ARRIERE, COLON, COMTE, COTE, FERME, FOSSE, JEUNE, MARCHE, MODELE, PASSE, PECHE, SECRETE, TRAITE, les formes ambiguës de REPARTIR,

et nous avons placé certaines combinaisons sûres dans notre lexique. Exemples:

UN FILM DOUBLE	(doublé)
MARCHE DE DUPES	(marché)
FONDE DE POUVOIR	(fondé)
JAMBON SALE	(salé)
SOUFFLE AUX FRAISES	(soufflé)
LES CHARGES DE FONCTION	(chargés)

Pour les autres cas, nous tentons de choisir la forme la plus fréquente. Exemples:

CHAQUE LETTRE	(*lettre*, plutôt que *lettré*)
IL REPARAIT	(*réparait*, plutôt que *reparaît*)
IL HAIT	(*hait*, plutôt que *haït*)
UN CAS ILLUSTRE	(*illustre*, plutôt que *illustré*)
IL DORT OU IL TRAVAILLE	(*ou*, plutôt que *où*)

Dans le programme L3, nous avons supprimé les analyses prévues pour les hétérophones qui appartiennent à la même catégorie grammaticale (FILS, JET, LACS, PUB, etc.) puisqu'elles sont inutiles dans L3; nous avons ajouté une analyse pour CHASSE (*chasse/châsse*), TACHE (*tache/tâche*), et les contrastes verbaux du type *fit/fît, mourut/mourût.*

Conclusion

Les tests récents auxquels nous avons soumis les programmes L1, L2, et L3 indiquent que nous avons pratiquement atteint le maximum des performances que l'on peut attendre d'une analyse linéaire. Le dernier test (d'environ deux mille mots) donne les résultats suivants:

1. Mots étrangers: Ce test contient un article de journal avec deux noms propres étrangers qui ne sont pas prononcés de façon acceptable.
2. Ambiguïtés linguistiques: Ce test contient plusieurs ambiguïtés linguistiques (*le pouvoir*, *en couronne*, UN COUTEAU A DEUX LAMES, EXCUSE DE LA FAUTE, etc.) qui ne peuvent pas être résolues sans analyse contextuelle. Le programme L1 fait trois analyses qui sont acceptables du point de vue linguistique, mais qui ne correspondent pas à la réalité sémantique (ex: *le pouvoir* est interprété comme 'pronom + infinitif' alors que le sens est 'déterminatif + nom'); pour les programmes L2 et L3, le nombre passe à 5 (ex. LE COUTEAU A DEUX LAMES est interprété comme *le couteau a deux lames* alors que le sens est *le couteau à deux lames*).
3. Erreurs d'analyse: Dans ce test, notre programme n'a fait aucune erreur d'analyse, c'est-à-dire que:
   a. les règles qui ont été appliquées étaient correctes,
   b. dans les cas où un mot ambigu a traversé le module d'analyse sans rencontrer de règles qui puissent s'appliquer à ce mot, l'attribution linguistique par défaut était correcte,
   c. aucun signal linguistique ne s'est révélé trop éloigné.

   Cela ne signifie pas que tous les tests futurs seront sans erreurs d'analyse. Il y aura certainement des tests où le signal linguistique sera trop éloigné et où des phrases ne seront pas analysées correctement par notre module; dans ce dernier cas, les règles nécessaires seront ajoutées.

Les résultats de ce test correspondent aux résultats obtenus par d'autres tests, c'est-à-dire une demi-douzaine de décisions critiquables pour un test d'environ 2000 mots. C'est une proportion extrêmement faible, mais il est vrai qu'une personne francophone cultivée aurait prononcé les deux mots étrangers de façon acceptable et aurait pu résoudre les cas d'ambiguïté linguistique. Est-il encore possible de réduire cet écart?

1. Mots étrangers:

N'importe quel mot de n'importe quelle langue peut apparaître dans un texte français, en particulier dans les journaux, magazines, et récits de voyages. Accidentellement, nos règles de phonétisation du français donnent une prononciation acceptable (*lockstep*, *Frühling*, *difficoltà*, etc.), mais, dans la plupart des cas, la prononciation est inacceptable.

Une personne cultivée — et précisément parce que c'est une preuve de culture — sait comment elle doit prononcer dans son milieu social un grand nombre de mots et noms propres étrangers. De plus, une personne cultivée qui rencontre pour la première fois un mot étranger peut généralement en déceler l'origine linguistique (par le contexte ou l'aspect général du mot) et en donner une prononciation acceptable ou — au moins — qui ne soit pas risible.

Il est tentant d'essayer d'établir un programme qui pourrait imiter la performance d'une personne cultivée, mais trois obstacles se présentent:

a. Comment déterminer qu'un mot ne doit pas passer par nos règles de phonétisation? Il serait facile de déterminer que *keepsake, oats, Kronprinz, avvocato, Deng Xiaoping, Mickiewicz* n'appartiennent pas au domaine français, mais que faire pour l'anglais *paper*, l'allemand *Brot*, l'italien *cambiare*, etc. qui ont des combinaisons de lettres acceptables en français? Pour être certain de déceler tous les mots qui n'appartiennent pas au domaine français, il faudrait que notre lexique contiennent toutes les formes de tous les mots et noms propres du français et que, par soustraction, tous les mots et noms propres non compris dans ce lexique soient considérés étrangers.

b. Si un mot a été marqué 'étranger', comment peut-on déterminer son origine linguistique? Les combinaisons de caractères qui n'appartiennent qu'à une seule langue sont rares. Ecrire un programme qui puisse interpréter le contexte et l'aspect général du mot ne semble pas réalisable.

c. Il faudrait ensuite faire passer le mot par des règles qui en donnent une phonétisation acceptable. Si le mot est espagnol, italien ou allemand, les difficultés ne seraient pas trop grandes, mais pour l'anglais — où les rapports entre les formes écrites et orales sont aussi complexes qu'en français — il faudrait un ensemble de règles de phonétisation et un lexique d'exceptions; les hétérophones (*read, lead, wind, row, lives*, etc.) exigeraient une analyse linguistique, analyse impraticable pour un mot anglais isolé dans un texte français.

Ces obstacles ne semblent pas avoir de solutions et comme il est impossible de mettre dans notre lexique des millions de mots anglais, allemands, russes, espagnols, etc., nous avons adopté la solution suivante:

a. Nous avons placé dans notre lexique plusieurs centaines de mots d'origine étrangère (ainsi que des mots bretons ou alsaciens) que les règles de phonétisation du français ne prononcent pas d'une façon satisfaisante. Exemples:

> Auschwitz, baby sitter, bagad, Beethoven, bowling, breeder, brushing, edelweiss, Goethe, Guebwiller, Heidelberg, Hemingway, hooligan, Huelgoat, huerta, in absentia, in extremis, jeans, kiwi, kleenex, kugelhof, meeting, outlaw, Paderewski, pretium doloris, Rubens, Schubert, Schwartz, Shakespeare, shakespearien, Shaw, sit-in, skin, skinhead, strip-tease, teen-ager, twist, Wagner, wehrmacht, zapateado, etc., etc.

b. L'utilisateur du programme peut modifier la prononciation des mots contenus dans le lexique et peut ajouter autant de mots qu'il le désire avec la prononciation qu'il préfère.

2. Position du signal linguistique:

Nous avons vu que, dans de rares cas, le signal linguistique est hors de portée de notre analyse linéaire et que l'analyse risque d'être fausse; par exemple, dans une des phrases de la page 10, *somnolent* est marqué adjectif alors que c'est un verbe. Dans ces phrases exceptionnelles, la position du signal linguistique dans le contexte gauche ou le contexte droit est imprévisible et les syntagmes qui se trouvent entre ce signal et le mot en cours d'analyse sont aussi imprévisibles: *qui aiment tant* pourrait être *qui disent ne plus aimer, qui font semblant d'aimer plus que jamais*, etc. Ces phrases ne poseraient aucun problème à un enfant d'une quinzaine d'années ayant reçu une formation scolaire normale, mais il est clair que, dans de tels cas, l'analyse linéaire risque d'échouer. Nous ne connaissons pas de système informatique qui, à l'heure actuelle, soit capable d'analyser de telles phrases avec 100% de succès.

3. Ambiguïtés linguistiques:

Lorsqu'une ambiguïté linguistique ne peut pas être résolue par notre analyse linguistique ou notre modeste analyse contextuelle, notre programme fait un choix basé sur des critères de fréquence; par exemple,

lorsque le texte ne contient pas de signal linguistique qui permette de décider si *boucher* dans la chaîne *sans le boucher* est un nom ou un infinitif, notre programme décide que *boucher* est un nom.

Dans notre programme, nous avons deux catégories pour les 48 homophones qui, comme *boucher*, peuvent être noms ou infinitifs; une catégorie contient 30 mots qui se rencontrent plus souvent comme infinitif que comme nom (ex: *avoir, toucher*) et l'autre contient 18 mots qui ont la fréquence contraire (ex: *clocher, officier*). Dans notre programme actuel, les mots de chaque catégorie sont traités de la même façon quelle que soit la construction. Nos tests ont montré que cette séparation en deux groupes n'était pas pleinement satisfaisante et que, dans certains cas, il aurait été utile de considérer chacun de ces 48 homophones séparément afin que l'analyse linguistique puisse spécifier que, dans la construction X, 33 de ces 48 mots sont des noms et que, dans la construction Y, 9 sont des infinitifs et que, dans la construction Z, ils sont tous des noms, etc.

De même, il nous a paru au début de ce travail qu'il était logique de placer *dit* et *fait* dans le même ensemble linguistique puisque ces deux mots peuvent être verbe, participe passé, et nom; l'expérience a montré qu'il aurait été préférable de les séparer car, dans certaines constructions ambiguës, la probabilité 'nom' est beaucoup plus élevée pour *fait* que pour *dit*.

Notre programme divise les prépositions en 18 groupes; nous séparons, par exemple, celles qui ne peuvent être suivies que d'un infinitif de celles qui ne peuvent être suivies que d'un nom ou adjectif; pour les liaisons, nous séparons les prépositions monosyllabiques des autres. Là encore, il aurait été préférable de pouvoir adresser chaque préposition séparément; par exemple, la préposition *comme* peut être suivie d'un pronom sujet ou d'un pronom objet:

Tu parles comme il/elle parle.
Tu parles comme lui/elle.

tandis que les autres prépositions ne peuvent être suivies que d'un pronom objet (*sans lui, pour eux, avec lui*, etc.).

Il s'agirait donc de remplacer des critères de groupe par des critères individuels; par exemple, dans le programme L1, les 682 mots du type *offre* (verbe/nom féminin), au lieu d'être tous soumis aux mêmes critères, seraient traités individuellement. Etablir ces statistiques individuelles de fréquence exigerait des milliers d'heures de travail car il faudrait déterminer, pour chacun des milliers de mots qui appartiennent à plusieurs catégories grammaticales, quel est, pour chacune des constructions ambiguës où ce mot peut se rencontrer, l'emploi le plus fréquent. Se lancer dans un tel travail pour essayer de modifier une demi-douzaine d'analyses critiquables dans un texte de 2000 mots ne semble pas justifié aujourd'hui, d'autant plus que nous ne pouvons pas être certains que ce travail serait nettement positif. Il faudrait, pour qu'il le soit, que la plupart des règles découlant de cette recherche nous donnent des rapports au moins égaux à 80% contre 20%; apprendre que *pouvoir*, dans la construction X, est un nom dans 55% des phrases testées (contre 45% pour l'infinitif) ne nous permettrait pas d'améliorer notre analyse.

Nous pouvons donc considérer que notre travail est terminé puisque nous avons réduit autant que nous le pouvions l'écart qui sépare ce que peut faire une machine de ce que fait une personne cultivée. Quelle que soit la méthode employée pour l'analyse linguistique du français, il apparaît douteux que cet écart puisse être réduit à zéro.

Ce constat n'est en rien négatif; nous voulons, au contraire, souligner le fait que, en ce qui concerne l'analyse linguistique, la performance de la machine peut être presque l'égal de la performance humaine. Mais il est tout aussi important de ne pas contribuer à éveiller ou à maintenir de fausses espérances quant au traitement général des langues humaines par l'informatique. Sans doute, les recherches qui sont en cours par un grand nombre de chercheurs dans des domaines complexes tels que la traduction automatique et la reconnaissance de la parole (domaines qui, eux aussi, dépendent d'une bonne analyse linguistique) sont nécessaires et seront utiles si leurs applications ne dépassent pas le cadre où elles sont performantes et

n'engendrent pas un phénomène de rejet qui nuirait aux linguistes tout comme aux informaticiens. Car, en effet, il n'est guère vraisemblable qu'un jour une machine puisse lire un texte en exprimant des sentiments de jalousie, d'amour, de haine, de joie, ou de tristesse, qu'elle puisse traduire automatiquement et sans erreurs les oeuvres de Proust, qu'elle puisse transcrire en orthographe standard et sans fautes d'accord une conversation entre deux interlocuteurs qui s'expriment sans modifier leur élocution habituelle, qu'elle puisse converser avec une autre machine comme deux êtres humains le font sur les sujets les plus divers, qu'elle puisse déchiffrer aussi bien que nous le faisons les textes écrits à la main, qu'elle puisse décrire oralement et par écrit une photographie ou une scène de la rue, qu'elle puisse analyser la prononciation d'un étudiant étranger et suggérer des corrections, qu'elle puisse examiner un texte écrit et corriger les fautes d'orthographe, d'accord, et de syntaxe, ou qu'elle puisse analyser cet article et en faire la critique.