AMERINDIA n°3, 1978

TEXTECONCORD
Programme de concordance générale

Guy BUCHHOLTZER*

Department of Anthropology and Sociology
University of British Columbia

A. INTRODUCTION

J'avais annoncé précédemment[1] la mise au point d'un programme de concordance générale destiné à faciliter la recherche sur les langues amérindiennes.

Après avoir appliqué ce programme à divers textes, aussi bien anglais et français que kwakwala ou okanagan (langues parlées respectivement par des Indiens kwakiutl et salish de la Colombie Britannique, Canada) et testé les résultats obtenus, nous nous sommes vu contraints, devant la complexité de la matière à traiter (tels les différents problèmes inhérents à la translittération[2], à l'homographie, aux différentes recherches contextuelles, etc.), de réécrire complètement ce programme et à le rendre ainsi apte à traiter plus efficacement des problèmes d'ordre linguistique.

---

* ERA 431 (C.N.R.S.) - Ethnolinguistique amérindienne.

1 Cf. Amerindia 1.

2 Translittération : Le texte est tout d'abord obtenu par la *transcription* des données orales fournies par l'informateur. Ce texte transcrit est ensuite *translittéré :* on transpose l'alphabet du texte transcrit dans un autre alphabet, arbitraire, compatible avec le langage utilisé par le système de l'ordinateur.

Le principe qui m'a guidé tout au long de ce travail a été à la fois celui de la *simplicité* et de la *souplesse* dans le choix des procédures informatiques et des différentes options offertes à l'utilisateur. Une autre priorité retenue a été *l'abaissement du coût de l'exploitation du programme.*

Mais d'abord, qu'est-ce qu'une concordance ? Simplement définie, c'est une liste de termes (ou de mots) généralement classés alphabétiquement et dont chacun est accompagné d'un contexte déterminé. Le but immédiat de la concordance est de fournir au chercheur intéressé tous les contextes possibles qui peuvent co-occurrer avec un terme donné d'un texte. Une concordance présente donc en quelque sorte tous les caractères d'un ouvrage de référence. Le meilleur emploi qu'on puisse faire d'un ouvrage de référence dépend – outre des qualités qui lui sont propres – du chercheur lui-même.

Les premières concordances connues datent environ du XIIe ou XIIIe siècle. Elles ont été toutes l'œuvre des exégètes et des littérateurs du Moyen Age. Citons pour mémoire les *Concordantiae bibliorum* de Conradus le Jeune, de Halberstadt (1480)[3]. Toutes ces concordances sont en langue latine. A partir du XVIe siècle apparaissent des concordances en d'autres langues. En Espagne, Domingo de Baltanus Mejía[4] fait paraître en 1555 les *Concordancias de muchos passos difficiles de la divina historia.* En Angleterre, c'est Robert Allen[5] qui en 1612 écrit une concordance des proverbes du Roi Salomon. Mais ce n'est qu'à partir de 1787, avec la parution à Londres d'une concordance des oeuvres de Shakespeare[6] que cette technique se trouve utilisée dans le domaine de la littérature non religieuse. Depuis et malgré la somme de travail proprement immense que représente ce genre d'entreprise – de l'ordre de plusieurs années, si ce n'est de plusieurs dizaines – de

---

3 Conradus le Jeune : de Halberstadt
Concordantiae bibliorum, Ed. Johan Otmar, Reutlingen, (756 p.), 1480.
Concordantiae bibliorum, Ed. Anton Koberger, Nuremberg, (792 p.), 1485.
Concordantiae bibliorum, Ed. Peter Drach, Spire (706 p.), 1485.

4 Domingo de Baltanus Mejía : Concordancias de muchos passos difficiles de la divina historia, Sevilla, Espania, Ed. Matin de Montesdoca, 1555.

5 ALLEN, Robert, *Concordances of the Holy Proverbs of King Solomon*, Ed. W. Hall & I. Beale for Welvie, London, 1612.

6 BECKET, Andrew, *A Concordance to Shakespeare*, Ed. G.G.J. & J. Robinson, London, (470 p.), 1787. Mais la concordance la plus connue des oeuvres de Shakespeare demeure sans doute celle de John BARTLETT, publiée à Londres en 1894.

nombreuses concordances ont paru[7]. Mais on peut dire avec raison que ce n'est qu'avec l'avènement des calculatrices électroniques que le phénomène a connu un réel essor et un regain d'intérêt à la fois pour les linguistes, les philologues, les historiens, les chercheurs en sciences sociales, les administrations, etc., tous ceux enfin qui travaillent sur des données textuelles. L'avantage que représente l'emploi de l'ordinateur réside dans l'exactitude, la rapidité et la possibilité de produire rentablement différents types de sortie selon les besoins spécifiques du chercheur.

La production de concordances est devenue réellement importante à partir des années 1970. Récemment cette technique a été introduite plus systématiquement dans la recherche sur les littératures non européennes[8].

Une concordance comporte généralement une liste alphabétique des formes lexicales, des tableaux fréquentiels et de co-occurrence. "Pour deviner l'âme du poète, ou du moins sa principale préoccupation, cherchons dans ses oeuvres quel est le mot ou quels sont les mots qui s'y représentent avec le plus de fréquence. Le mot traduira l'obsession" disait Baudelaire.

Mais de fait, une concordance est bien plus qu'une liste alphabétique de mots ou qu'un tableau fréquentiel sur le vocabulaire de l'auteur.

C'est tout d'abord un instrument de recherche linguistique proprement dit. Un programme de concordance doit pouvoir traiter aussi bien d'unités lexicales simples que complexes. Par exemple, "chemin", "chemin de fer" et "ferro-nickel" doivent être distingués.

De même, si l'analyse est conduite au niveau phonologique ou morphologique, les distinctions nécessaires à la différenciation des unités linguistiques alors mises en jeu doivent être rendues possibles durant l'utilisation du programme. En d'autres termes, le programme doit être apte à détecter les limites données par le chercheur lui-même aux unités linguistiques qu'il traite. Ces unités linguistiques peuvent être aussi bien des unités lexicales simples que

---

7 VINOGRADOV, V.V. *Slovar' jazyka Pushkina*, Akademia Nauk SSSR, Moskva, 1956. C'est un ouvrage de 4 volumes dont l'établissement a demandé plus de 20 ans de temps. Quant aux concordances de la Vulgate, du XIIIe siècle, plus de cinq cents dominicains y auraient travaillé durant plus d'un siècle. Les concordances de la Vulgate, du XIIe siècle, auraient demandé la participation de plus de cinq cents dominicains.

8 Bien qu'il existe une concordance du Coran établie au XVIIIe siècle, les recherches les plus actives concernent surtout 1'Extrême-Orient. Citons, pour mémoire, *A concordance to the poems of Li Ho (790-816),* réalisée par R.C. Irick (1969). Cette concordance est en langue chinoise. D'autres concordances ont été réalisées depuis lors sur des textes chinois et japonais.

composées ("mot", "lexie" ; "lexie composée"[9]), des morphèmes ou des phonèmes, des phrases ou des énoncés entiers, etc. Ceci pose le problème de la *définition du contexte*. En général, les programmes de concordance ne fournissent qu'un contexte défini *mécaniquement* (tant de mots à gauche, et tant de mots à droite du mot concordé, par exemple) et ne traitent que d'unités lexicales telles que les "mots" séparés les uns des autres par des "blancs" ou par des signes de ponctuation éventuels. Grâce aux sous-programmes "Alphabet", "Recherche contextuelle" et au programme "Contexte", TEXTECONCORD possède une "capacité d'analyse linguistique" qui lui est propre.

TEXTECONCORD permet à l'utilisateur de déterminer lui-même, le type d'alphabet utilisé étant entendu qu'une lettre donnée de cet alphabet peut être composée de plusieurs caractères à la fois. (unigraphe, digraphe, trigraphe, etc. : une lettre de l'alphabet choisi pouvant contenir jusqu'à dix caractères) (cf.§ B.2,3). L'utilisateur choisit le type de contexte qu'il désire en fonction donc de ce qui vient d'être dit plus haut au sujet des unités linguistiques analysées. (Cf. les différentes procédures § B.5). Par ailleurs, il a le choix entre deux options fondamentales (cf. § B.6) :

a) Production d'une concordance de toutes les occurrences de tous les mots du texte selon un ordre alphabétique donné et selon leur ordre d'occurrence dans le texte.

b) Production d'une concordance de toutes les occurrences de tous les mots du texte selon un ordre alphabétique donné et quelque soit l'ordre d'occurrence dans le texte.

Dans le cas de (b), il est possible de classer *l'ensemble*, (et ceci jusqu'à une profondeur d'une quinzaine d'unités linguistique – phonèmes, morphèmes ou lexies etc. –) des contextes selon l'ordre alphabétique – ou alphanumérique – choisi. Grâce à cette capacité, la production par TEXTECONCORD d'une concordance, ainsi définie, permet d'envisager des études très serrées sur des problèmes précis de phonologie et de syntaxe.

Une dernière propriété fondamentale réside dans les choix offerts au chercheur dans la *définition des contextes* à l'aide du sous-programme "Recherche contextuelle" (cf. ci-dessus et § B.8).

---

[9] Cf. Bernard POTTIER, *Linguistique générale* ; 1ère partie, Ed. Klincksieck, Paris, 1974.

Notons que, si besoin est, il est possible aussi de recourir au programme CONTEXT. Celui-ci possède la particularité de *générer* des *suites* d'unités linguistiques à partir du contenu de listes d'unités linguistiques (phonèmes, morphèmes, lexies, lexies composées, etc.) préalablement données par le chercheur. Le programme génère à partir du contenu de ces listes, par l'intermédiaire de règles booléennes, *l'ensemble* des suites, "phrases" ou combinaisons phonologiques désirées etc., et compare ces dernières à celles qui existent *effectivement* dans le texte, que ce dernier soit le texte original ou qu'il soit le texte produit par la concordance. Il est alors éventuellement possible de n'éditer que les suites lexicales du texte compatibles avec celles définies par le chercheur, ou de ne retenir que le texte qui ne les contient pas. Il est possible d'adjoindre et de faire correspondre à chacune de ces unités linguistiques – choisies par le chercheur, répétons-le – un ensemble de termes descriptifs (caractéristiques syntaxiques ; propriétés sémantiques ; caractéristiques de la distribution dans le mot, dans la phrase, etc. ; modalités, notations paraphrastiques, indications contextuelles, distributionnelles, extralinguistiques – contextes ethnologiques, conditions d'énonciation, etc.). On voit là l'intérêt d'une telle procédure pour mener à bien une recherche approfondie sur un aspect donné d'une langue ou sur un problème spécifique de la communication linguistique ; ou tout simplement pour résoudre le problème de l'homonymie, de l'homographie, et pour regrouper en sortie les différentes réalisations d'un même verbe, ou encore pour produire des lexiques particuliers ou des champs sémantiques, ou répertorier les catégories conceptuo-grammaticales etc.). Elle permet également de produire une mise en correspondance lexicale entre les termes ou groupes de termes de deux ou trois langues différentes (traduction, lemmatisation).

Le lecteur-chercheur a également le choix d'éliminer de la concordance tous les mots qu'il ne désire pas avoir en sortie ; ou de n'en donner qu'un index, de ne concorder qu'une suite de mots donnée, ou de ne travailler que sur des variantes. Le lecteur définit lui-même tous les types de sortie possibles (contexte limité à la phrase, indexes, etc.) et le format de présentation qu'il désire qu'ils aient (cf. § B.10).

En résumé, TEXTECONCORD peut produire plus d'une vingtaine de concordances différentes.

J'ai tenu à ce que le programme possède de par lui-même la capacité de conduire l'ensemble de ces analyses sans qu'il soit nécessaire pour autant de

passer par un codage préalable du texte. Ce précodage est une opération d'un *coût extrêmement élevé* et pose de délicats problèmes de *contrôle de données*. Grâce à un système de "clés-options" l'usager définit le type de travail qu'il veut conduire. Le texte doit être bien entendu perforé sur cartes (ou enregistré sur bandes magnétiques) ; chaque ligne du texte comportant un numéro (identificateur). Un programme de mise en correspondance "TRANSCOD" permet éventuellement de retranscrire les résultats sous leur forme graphique originale.

Le chercheur soucieux de planifier sa recherche aura avantage à utiliser le sous-programme "STATUS" avant d'entreprendre une concordance d'un type donné sur un texte important afin d'en connaître le coût éventuel (cf. § B.1 pour une estimation), bien que celui-ci soit dans l'ensemble relativement bas.

Il est possible d'obtenir également un tableau des fréquences d'occurrences des différentes unités linguistiques qui constituent le texte. L'ensemble de ces procédures permet de réaliser plus systématiquement une analyse du contenu du texte.

B. DESCRIPTION GÉNÉRALE

TEXTECONCORD est un programme informatique permettant de produire une concordance générale de toutes les occurrences des mots de textes qui peuvent appartenir à n'importe quelle langue ou dialecte et quel que soit le système de transcription utilisé.

A TEXTECONCORD est adjoint une série de programmes et de sous-programmes. Ce sont :

- le programme MIXCORD : il permet d'obtenir une concordance générale à partir de deux autres concordances (mélange de deux concordances).

- le programme TRANSCORD : retranscrit sous leur forme originale les caractères translittérés du texte.

- le sous-programme STATUS : produit une statistique relative au coût de revient (en temps-machine, en coût réel) de la concordance en fonction du type et de la masse de données à traiter.

- le sous-programme MOTCOUNT : fournit une statistique sur les données (pas exemple, décompte des unités linguistiques, pourcentages, probabilités, etc.).

TEXTECONCORD a été tout spécialement conçu pour assurer une très grande souplesse d'utilisation dans des domaines d'application différents ; nous avons évité les trop nombreuses contraintes et la trop grande spécificité des programmes de concordance actuellement existants.

TEXTECONCORD permet à tout utilisateur de définir lui-même, au moyen de simples paramètres (cf. "Mode d'utilisation du programme de concordance TEXTECONCORD") :

- les types de données et leur mode d'indexation éventuel
- le type de concordance désiré
- le type de contexte voulu en sortie de traitement.

C. TYPES DE DONNÉES ET INDEXATION

a) Types de données :

Les données peuvent provenir de textes écrits dans n'importe quelle langue. Il est possible de distinguer trois types de textes :

- les textes écrits avec des caractères non latins (par exemple le chinois, l'hébreu, le japonais, le russe, etc.). Ces textes doivent être d'abord transcrits et translittérés en des caractères alphanumériques compatibles avec le système informatique utilisé (cf. en Annexe (1) la liste des caractères à utiliser pour IBM-370 et CDC-6600).

- les textes originalement écrits en caractères latins (par exemple, les textes de langue anglaise, française, espagnole, etc.).

- les textes provenant de la transcription de langues non écrites (par exemple, les langues amérindiennes, africaines, océaniennes, etc.).

La translittération de ces textes, souvent obligatoire avant leur traitement par ordinateur, nécessite parfois l'emploi de *plusieurs* caractères

alphanumériques pour une seule et même lettre de l'alphabet du texte original**. En espagnol par exemple, la prépalatale **č** peut être translittérée par **C*** (deux caractères) ; la nasale palatalisée **ñ** (dans "español") par **N+** (deux caractères). Dans ce cas, "español" sera translittéré par **ESPAN+OL**. Dans les langues amérindiennes, la translittération des phonèmes nécessite parfois l'emploi de trois ou quatre caractères. Exemple la vélaire labialisée **g^{w}** en kwakwala (groupe Wakash du nord-ouest du Canada) sera translittérée par **G_W** par exemple. **G_W** (trois caractères) constituera une seule lettre de l'alphabet translittéré du kwakwala.

TEXTECONCORD établit, en version "standard", une correspondance de un à un entre les lettres, de l'alphabet du texte original et celles de l'alphabet translittéré. Dans la version "non-standard", cette correspondance peut aller de un à cinq : le programme traite un groupe de caractères comme constituant une seule lettre de l'alphabet translittéré ; le nombre maximum de caractères est ici de cinq, mais il peut être étendu exceptionnellement à dix (dix caractères seront traités comme formant une seule lettre).

Cette possibilité permet également de reclasser des unités linguistiques (phonèmes, morphèmes, parties de mots, groupes de caractères, etc.) dans le cas d'études de la morphologie d'une langue, de recherches linguistiques comparatives, etc., et, aussi, dans l'étude approfondie de textes à usages administratifs ou juridiques.

Autres caractéristiques

Le type de données se définit également par le type de *ponctuation* utilisé. L'utilisateur de TEXTECONCORD devra indiquer les caractères (alphanumériques, diacritiques) réservés à la ponctuation du texte.

De même, il indiquera les caractères qu'il ne désire pas *classer* (par exemple : l'apostrophe ', les parenthèses, etc.), c'est-à-dire, les caractères alphanumériques qui ne rentrent pas dans l'alphabet translittéré choisi. Ces caractères non classés figurent, à leur bonne place, en sortie de traitement.

** Nous considérons l''alphabet comme un ensemble *fini* d'éléments, sans répétition. Un élément peut être représenté par un seul caractère (Ex. **a**, **b**, **y**, etc.) ou par plusieurs (Ex. : **A!**, **G"**, **L/**, etc.). Cet élément constitue la *lettre* de l''alphabet considéré.

b) Support des données

TEXTECONCORD traite les données qui ont été mises sur les supports matériels suivants : cartes perforées, disques et bandes magnétiques.

- cartes perforées : chaque carte perforée comprend quatre-vingts colonnes. Chacune des colonnes correspond à un seul caractère du texte translittéré, blancs (séparateur de deux mots) et signes de ponctuation compris. Chaque carte correspond à *une* seule ligne de données pour l'ordinateur ; une ligne de texte peut, bien entendu, comprendre autant de cartes perforées (c'est-à-dire, autant de lignes de données pour ordinateur de 80 caractères maximum chacune) qu'il est nécessaire; tout dépend de sa longueur.

- disques : chaque ligne du disque comprend au maximum 255 caractères chacune, sur IBM 370.

- bandes magnétiques : l'information que supporte la bande magnétique est divisée en *blocs* dont chacun supporte un à plusieurs enregistrements logiques (unités logiques de données) qui comprennent chacun à leur tour un certain nombre de caractères (bytes). L'utilisateur de TEXTECONCORD a la faculté d'enregistrer ses données, provenant de cartes perforées ou de disques, sur ces bandes magnétiques et d'appliquer à celles-ci le programme de concordance.

c) Indexation des données

On indexe généralement les données textuelles. Par exemple, à chaque ligne de texte correspond un numéro. Ou bien, on applique au texte une classification plus élaborée. On divise, par exemple, le texte en : numéro de volume, numéro du chapitre, numéro de la page, numéro de la ligne. Cette indexation constitue l'*identification* du texte et l'ensemble des chiffres qui caractérisent une ligne de texte constitue l'*identificateur*. Celui-ci permet de retrouver chaque mot à la fois dans le texte et dans la concordance.

Exemples d'indexation

	VOL	PAGE	COL.	LIGNE
COMPUTABLE (K M'PJU:T BL),A. CALCULABLE	002	C44	002	001
◄——— texte ———►		identificateur		

correspond au texte translittéré contenu dans le dictionnaire *Harrap's Shorter French and English Dictionary*, Ed. Bordas, Paris 1967. L'identificateur réfère à :

002 deuxième partie du volume (partie 'English-French')
C44 correspond à la page C:44 du volume, deuxième partie
002 deuxième colonne de texte
001 première ligne de texte

		LIVRE	POESIE	PAGE	LIGNE
2.	DE VELOURS ET D'ORANGE LA MAISON SENSEE	002	005	110	016
	D'ARGENT DETRUIT DE CUIR DE PLANCHES	002	005	110	017
	LA MAISON ACCUEILLANTE	002	005	110	018

Ceci correspond aux trois premiers vers de la poésie "Les excellents moments" de P. Eluard, dédiée à Francis Poulenc (en 1942). Cette poésie fait partie du Livre Ouvert (*livre 2*), *cinquième* poésie, page *110*, lignes *16* à *18* de l'édition NRF-Gallimard, Paris, 1947.

L'utilisateur de TEXTECONCORD peut librement décider du type d'identificateur désiré. L'utilisateur précisera :

- le nombre *maximum* de caractères (blancs et identificateur compris) d'une ligne de données indexée.

- la région où prend place l'identificateur.

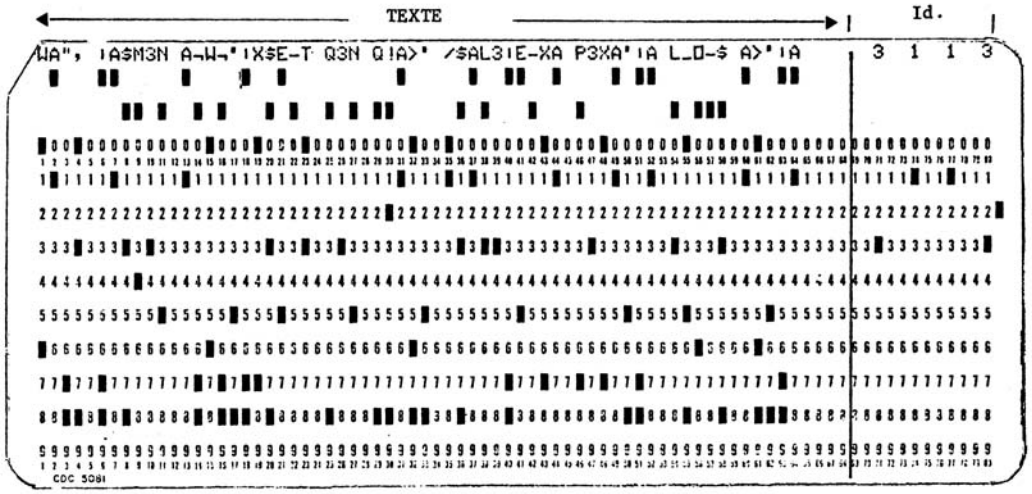
- le nombre de caractères réservés à l'identificateur.

TEXTECONCORD comporte deux options concernant l'identification du texte :

- la version "standard" : le texte est sur cartes perforées de 80 colonnes chacune. Le texte occupe les 68 premières colonnes ; l'identificateur les 12 dernières.

- la version "non-standard" : chaque ligne de texte peut comporter jusqu'à 255 caractères, identificateur compris. L'identificateur peut se trouver à n'importe quelle hauteur de la ligne du texte.

Exemple d'indexation "version standard" : carte perforée



Nombre de caractères utilisés par ligne : 80
Région du texte : 1 à 68
Région de l'identificateur : 69 à 80
Nombre de caractères réservés à l'identificateur : 12

Exemples d'indexation "version non-standard"

255 caractères maximum

identificateur		texte
ou		
texte		identificateur
ou		
texte	identificateur	texte

Il suffit d'indiquer à l'ordinateur le format d'entrée des données.

Ces deux versions (standard, non-standard) permettent de traiter n'importe quel type de texte indexé. On trouvera dans "Mode d'utilisation du programme de concordance TEXTECONCORD", la façon d'indiquer à l'ordinateur le type d'indexation choisi.

D. TYPE DE CONCORDANCE

Principe général

Le classement des occurrences en sortie peut se faire de deux façons :

a) - TEXTECONCORD produit une concordance de toutes les occurrences de tous les mots du texte selon leur ordre d'occurrence dans le texte.

L'ordre de sortie (O.S.), (alphabétique, alphanumérique, etc.) est donné par l'utilisateur.

b) - TEXTECONCORD produit une concordance de toutes les occurrences de tous les mots selon l'ordre de sortie donné par l'utilisateur et quel que soit l'ordre d'occurrence du mot dans le texte. Dans ce cas, tous les contextes sont classés selon l'ordre de sortie donné par l'utilisateur.

Pour chacune des options a. et b. ci-dessus, il est également possible de classer ensemble des mots graphiquement différents en donnant aux caractères constitutifs les mêmes équivalences numériques dans l'ordre de sortie O.S. De même, les concordances des types a. et b. peuvent fournir toute une série de résultats différents selon que la séquence des caractères constituant l'alphabet choisi comprend une ou plusieurs unités graphiques (caractères alphabétiques ou alphanumériques).

La concordance, de plus, peut être "ouverte" ou "fermée" (blocked, unblocked). En concordance ouverte, chaque occurrence de mot concordé est accompagnée d'une largeur arbitraire de texte défini par l'ordinateur lui-même. En concordance fermée, c'est l'utilisateur qui définit le type de contexte désiré en sortie.

Tout ce qui précède est vrai à la fois pour les versions "standard" et "non-standard".

En concordance ouverte, le type de sortie peut présenter l'aspect ci-dessous :

Exemple : (extrait de "Morphology of the Folktale" de V. Propp, 1973, published for *The American Folklore Society*)

TALES. BUT SUCH A REQUIREMENT IS BASED ON DELUSION. LET US DRAW	1	01	15	14
AN ANALOGY. IS IT POSSIBLE. TO SPEAK ABOUT THE LIFE OF A LANGUAGE	1	01	15	15
WITHOUT KNOWING ANYTHING ABOUT THE PARTS OF SPEECH, E.E., ABOUT	1	01	15	16
CERTAIN GROUPS OF WORDS ARRANGED ACCORDING TO ... etc...				

Type de sortie obtenue : (le contexte est en réalité plus large que nous le figurons ci-dessous)

TALES. BUT SUCH A REQUIREMENT IS	BASED	ON DELUSION. LET US DRAW	1	01	15	14
DRAW AN ANALOGY. IS IT POSSIBLE TO	SPEAK	ABOUT THE LIFE OF A	1	01	15	15
KNOWING ANYTHING THE PARTS OF	SPEECH,	I.E., ABOUT CERTAIN	1	01	15	16
contexte gauche	mot concordé	contexte droit	identificateur			

Caractéristiques principales des différents types de concordances

L'utilisateur définira lui-même pour tous les types de concordance qui suivent, l'*ordre de classement en sortie* de toutes les occurrences des mots du texte : cet ordre peut être alphabétique ou alphanumérique par exemple, ou autre. S'il le désire, des mots commençant par des caractères différents peuvent être groupés ensemble.

Choix de concordances que peut faire l'utilisateur de TEXTECONCORD :

1) obtenir une concordance de *toutes* les occurrences de *tous* les mots du texte selon leur ordre d'occurrence dans le texte et selon un ordre de sortie laissé à son choix (cf. § 6, a. & b.).

2) obtenir la concordance de *toutes* les occurrences dans le texte d'une liste de *mots donnés en entrée*.

3) concordance de *toutes* les occurrences de tous les mots du texte *exception* faite de certains mots. L'utilisateur donnera, en entrée, la liste des mots qu'il ne désire pas concorder.

4) concordance de *toutes* les occurrences de tous les mots qui suivent un caractère diacritique ou autre donné. Exemple : concordance de tous les mots qui débutent une phrase, c'est-à-dire ceux, qui suivent en principe un signe de ponctuation donné ; le point, par exemple. L'utilisateur donnera la liste des caractères diacritiques ou alphanumériques remplissant cette fonction de délimiteur.

5) concordance de toutes les occurrences de tous les mots ou groupes de mots qui se trouvent placés entre deux mêmes signes diacritiques donnés. L'utilisateur peut ainsi, concorder *à part* les mots, groupes de mots., paragraphes, etc., qu'il aura mis entre deux mêmes signes diacritiques donnés. Par exemple *MOT*, *la question administrative...* seront concordés à part.

6) Possibilité de *ne concorder que* les mots placés entre deux mêmes signes diacritiques donnés. Ces signes diacritiques sont appelés *marqueurs*. Le nombre maximum de marqueurs différents est de trois et ils sont donnés par l'utilisateur. Un marqueur peut, bien entendu, être constitué par une lettre (de *un* seul caractère) de l'alphabet du texte translittéré. Cette possibilité constitue une des fonctions extractive du programme.

7) Possibilité d'indiquer, dans la concordance, certains mots par leur index seulement (identificateur) ; ceci évite l'impression, longue et coûteuse, des contextes de certains termes dits "mots-outils" ou "mots-grammaticaux" (par exemple, en français : à, ce, de, et, le, la, etc.).

8) *Recherche contextuelle*

a: Concordance de toutes les occurrences de tous les mots du texte associés aux mots donnés dans une liste. Les mots de la liste peuvent précéder et/ou suivre le mot concordé.

Exemple : Soient les mots (ou parties de mots) a1, a2, a3, ... a^n donnés dans une liste A et MOT (1, 2, 3, ..., n) les occurrences de tous les mots concordés du texte. On aura, par exemple, en sortie :

----------a1	MOT1	----------	ID
----------	MOT2	a3----------	ID
……	……	……	
……	……	……	
----------a2	MOTn	a5----------	ID
	etc.		
contexte gauche		contexte droit	identificateur

b: Concordance des occurrences des mots donnés dans une liste A et associés avec des mots donnés dans une liste B. A et B peuvent aussi bien être constitués par des phonèmes, morphèmes ou toute autre unité lexicale ou grammaticale, etc.

Exemple : Soient A1, A2, ..., Ai, ..., An les mots de la liste A et B1, …, Bi, ..., Bn les mots de la liste B. Les mots A peuvent être ou non différents de ceux de B.

Exemple de sortie :

----------B2	A4	B8----------	ID
----------B1	A12	B34----------	ID
……	……	……...	
……	……	……..……..…	
----------B24	A3	----------	ID
----------	A7	B9----------	ID
	etc.		
contexte gauche		contexte droit	identificateur

Note importante : l'utilisateur de TEXTECONCORD qui fait cette recherche contextuelle peut avoir, s'il le désire, en même temps que les résultats qui précèdent, l'ensemble de la concordance de toutes les occurrences de tous les mots du texte.

c: Concordance de toutes les occurrences de tous les mots du texte qui sont *suivis* par une suite de mots (ou autres unités linguistiques) donnés dans une liste.

Exemple : Soit la liste des suites de mots suivants

1.	mot1 mot3 mot16
2.	mot5 mot4
…	……..
n	mot5 moti mot1

et MOTi (i = 1,n) les occurrences des mots concordés.

Exemple de résultat :

----------	MOT2	mot1	mot3	mot3---	ID
----------	MOT8	mot5	moti	mot1---	ID
……	……	……			
……	……	……			
----------	MOTn	mot5	mot4----------		ID
contexte gauche			contexte droit		identificateur

9) Concordance de toutes les occurrences de tous les mots avec la possibilité de grouper ensemble, en sortie, des mots appartenant à des listes de mots alphabétiquement différents.

Autres caractéristiques de TEXTECONCORD

L'ensemble des résultats peuvent se présenter sous la forme suivante : (il est possible de supprimer, un, ou l'ensemble, des trois premières possibilités)

1/

Titre général (5 lignes de 80 caractères chacune)

2/

Introduction ou commentaires de 50 lignes maximum (80 caractères par ligne)

3/

Texte à concorder (longueur illimitée)

4/

Concordance du texte

D'autre part :

a - toutes les pages de la concordance sont numérotées

b - les occurrences des mots commençant par un caractère différent sont séparées les uns des autres par un double interligne

c - chaque page de la concordance peut comporter un titre de quatre-vingts caractères maximum (blancs compris)

d - mis en vedette du mot concordé avec indication de son nombre réel d'occurrence dans le texte. Edition d'un index.

E. TYPES DE CONTEXTES

Nous avons vu qu'il existe, en ce qui concerne le contexte, deux possibilités : concordance ouverte, concordance fermée.

En principe, et quel que soit le type de concordance choisi; les seules limites imposées au contexte, sur une *même ligne d'impression*, sont dictées par la largeur du paravent (l'ordinateur imprime au maximum 132 caractères par ligne) et le choix de la séquence de caractères définissant l'alphabet choisi pour la translittération. L'utilisateur doit spécifier le type de contexte qu'il désire avoir en sortie.

Prenons quelques exemples de textes (extraits) : tous les caractères du texte original ont été translittérés ; les identificateurs sont arbitraires.

1) *Le Corbeau* d'Edgar Poe :

AND THE RAVEN, NEVER FLTTING, STILL IS SITTING, STILL IS SITTING	25	01
ON THE PALLID BUST OF PALLAS JUST ABOVE MY CHAMBER DOOR	25	02
AND HIS EYES HAVE ALL THE SEEMING OF A DEMON'S THAT IS DREAMING	25	03

etc...

2)

QUELLE DECEPTION, DEVANT LA PERVERSITE CONFERANT A JOUR	1	3	1
COMME A NUIT, CONTRADICTOIREMENT, DES TIMBRES OBSCUR ICI LA*CLAIR	1	3	2

Extrait de *variations sur un sujet* de Mallarmé (Divagations), Pléiade, p. 364.

3)

HERE, UNDER LEAVE OF BRUTUS AND THE REST,	1	21	045
FOR BRUTUS IS AN HONOURABLE MAN ;	1	21	046
SO ARE THEY ALL, ALL HONOURABLE MEN,	1	21	047
COME I TO SPEAK IN CAESAR'S FUNERAL.	1	21	048
HE WAS MY FRIEND, FAITHFUL AND JUST TO ME :	1	21	049
BUT BRUTUS SAYS HE WAS AMBITIOUS ;	1	21	050
AND BRUTUS IS AN HONOURABLE MAN.	1	21	051

Brutus, de Shakespeare

4) "Gexden" mythe kwakiutl récolté par F. Boas (*Kwakiutl Texts*, publications of the Jesup North Pacific Expedition, vol. III ; Leiden, 1902-1905)

Texte original :

ġō'kulaᵉlae ts!E'lqwalōlEla lā 'xa ëk̇ !eaᵉwī'nagwisa	1	1	3	21
hë'ẋ ᵉidā'Emlā'wis ġē'xdEn lā qae'sᵉid lāxa	1	1	3	22

Texte translittéré :

G*O-'KUlA$lAE TS!3'lQWAlO-/31A 1A- 'XA E"K*!E- A$WI-'NAGWI$A.	1	1	3	21
HE"'X*$IDA3M$lA-'WIS G*E-'XD3N lA- QA -'S$I-D lA-XA	1	1	3	22

5)	UN MILIEU ELEGANT EST CELUI OU" L'OPINION		89	9
	DE CHACUN EST FAITE DE L'OPINION DES AUTRES.		89	10
	EST-ELLE FAITE DU CONTREPIED DE L'OPINION		89	11
	DES AUTRES ? C'EST UN MILIEU LITTERAIRE.		89	12

Les plaisirs et les jours de Marcel PROUST, Ed. Gallimard, Paris 1924.

6)	SI CE MYTHE EST TRAGIQUE, C'EST QUE SON HEROS	4	163	27
	EST CONSCIENT. OU" SERAIT EN EFFET SA PEINE, SI A	4	163	28
	CHAQUE PAS L'ESPOIR DE REUSSIR LE SOUTENAIT?	4	163	29

(A. CAMUS, *Le mythe de Sisyphe*, 4ᵉ partie, p. 163, Ed. Idées-NRF/ Gallimard; Paris 1942).

Types de contextes en sortie

A. Contexte arbitraire (concordance ouverte), cf. supra.

B. Contextes définis par l'utilisateur (concordance fermée).

a) *Contexte gauche* : le mot concordé se trouve placé à droite de la page, suivi de l'identificateur.

Exemple 1.

STILL IS SITTING ON THE PALLID BUST OF PALLAS JUST ABOVE MY	CHAMBER	25	02
◄——— texte ———	mot concordé	Id.	

Exemple 2.

…..…….…….....…….QUELLE DECEPTION, DEVANT	LA	1	3	1
A NUIT, CONTRADICTOIREMENT, DES TIMBRES OBSCUR ICI,	LA*	1	3	1

b) *Contexte droit* : le mot concordé se trouve à gauche de la page, suivi du contexte.

Exemple 3

BRUTUS	AND THE REST, FOR BRUTUS IS AN HONOURABLE MAN; SO ARE THEY ALL	1	21	045
BRUTUS	IS AN HONOURABLE MAN; SO ARE THEY ALL, ALL HONOURABLE MEN,	1	21	046
BRUTUS	SAYS HE WAS AMBITIOUS; AND BRUTUS IS AN HONOURABLE MAN.	1	21	050
BRUTUS	IS AN HONOURABLE MAN	1	21	051

c) *Contexte défini* par l'étendue de la phrase ou de l'énoncé, ou du chapitre, etc. L'utilisateur peut faire figurer, seulement, le contexte compris entre deux mêmes caractères diacritiques donnés. Par exemple, si ces caractères sont constitués par le point (terminaison d'une phrase), le mot sera accompagné de son contexte comprenant le reste de la phrase.

Exemple tiré de César VALLEJO. Les rectangles [ ] représentent le contexte de la phrase.

[ ]. EL TRAJE QUE VESTI MAN# ANA. [ ] [ ] 21 3 27

Résultat :

	EL	TRAJE QUE VESTI MAN# ANA.	21	3	27
EL TRAJE QUE VESTI	MAN# ANA.		21	3	27
EL	TRAJE	QUE VESTI MAN# ANA.	21	3	27
EL TRAJE	QUE	VESTI MAN# ANA.	21	3	27
EL TRAJE QUE	VESTI	MAN# ANA.	21	3	27

mot concordé

d) *Contexte large* : chaque mot est accompagné d'un contexte plus large, de deux à trois lignes. Les contextes droits et gauches correspondant aux différents mots concordés sont séparés les uns des autres par un interligne.

Exemple (cf. texte supra, alinéa 3).

1. FOR BRUTUS IS AN HONOURABLE MAN:
   SO ARE THEY ALL, ALL HONOURABLE MEN,
   COME 121 048
   I TO SPEAK IN CAESAR'S FUNERAL
   HE WAS MY FRIEND, FAITHFUL AND JUST TO ME:
2. COME TO SPEAK IN CAESAR'S FUNERAL.
   HE WAS MY FRIEND,
   FAITHFUL 121 049
   AND JUST TO ME:
   BUT BRUTUS SAYS etc...

Le format de sortie est laissé au choix de l'utilisateur.

e) *Contexte étendu* : Par contexte étendu nous entendons aussi bien le contexte linguistique lui-même, (étude des structures paraphrastiques par exemple) que le contexte extralinguistique. Ce type de recherche implique l'utilisation de programmes dont nous donnons la description ailleurs.

1. Production d'une concordance réverse (importante pour l'étude des différents types d'affixes).

2. Analyse morphologique (production de lexiques principaux, de lexiques d'affixes et de racines) permettant de délimiter les classes syntaxiques et sémantiques.

3. Analyses contextuelles approfondies : programme CONTEXT.

4. Constitution d'un dictionnaire : a) unilingue b) bilingue.

5. Détermination des classes concepto-grammaticales.

6. Constitution d'index multi-entrées. Certains de ces index permettent d'opérer la mise en rapport entre données linguistiques et données anthropologiques (culture matérielle, organisation sociale, ethnohistoire, ethnobotanique, etc.), ainsi que des données muséologiques.

7. Etudes statistiques, études de corrélation.

8. Diverses analyses de contenu.

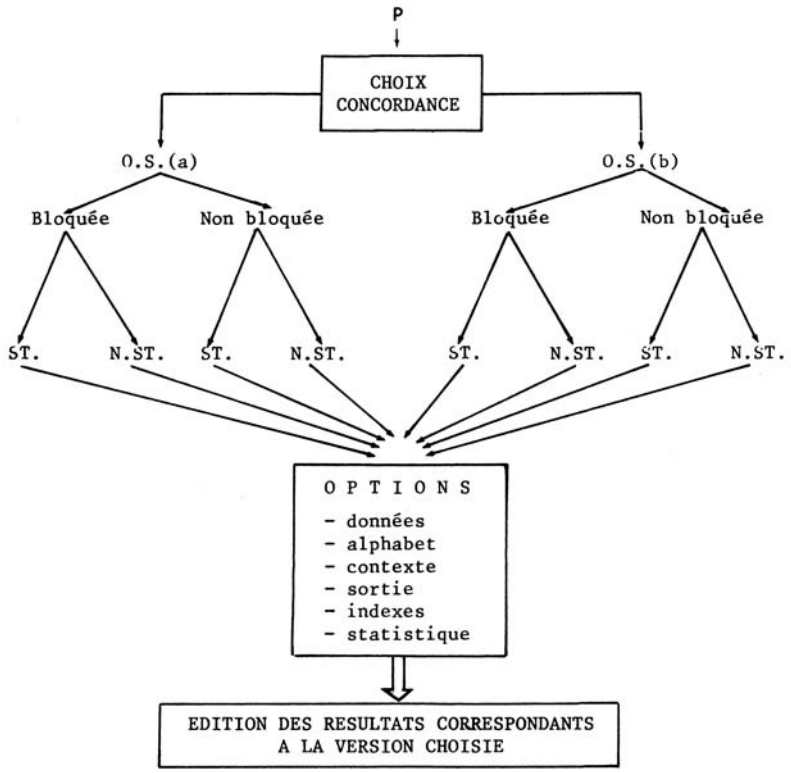
F. PROBLÈME DE L'HOMOGRAPHIE

Ce problème peut être résolu par le secours aux sous-programmes "recherche contextuelle" de TEXTECONCORD (cf. D, alinéas 8 et 9) et au programme CONTEXT (cf. E, § e, alinéa 3).

OPTIONS

Organisation du programme TEXTECONCORD : Choix du type de concordance :

Les différents types de concordance sont sélectionnés par l'utilisateur au moyen de *paramètres*. Ces derniers sont organisés selon une hiérarchie opératoire dont on a donné ci-dessous une représentation schématique.



Notes :

L'ordre de sortie O.S. des mots du texte traité est donné par l'utilisateur (alphabétique ou alphanumérique). De plus, la concordance se fait de deux manières différentes :

- O.S.(a) : occurrence de tous les mots selon leur ordre d'occurrence dans le texte ;
- O.S.(b) : occurrence de tous les mots quel que soit leur ordre d'occurrence dans le texte.

Le symbole ST. signifie version standard. Le symbole N.ST. : version non-standard.

Ce système permet de produire plus d'une vingtaine de types de concordance différents.

CONCLUSIONS

TEXTECONCORD est utilisable sur les ordinateurs de la série IBM-370/168 et CONTROL-DATA-6600.***

LANGAGE DE PROGRAMMATION UTILISE

Le langage qui a été utilisé pour écrire le programme TEXTECONCORD est le FORTRAN IV, version G.

Ce langage est à la fois compatible avec les systèmes d'IBM et de Control Data. La programmation a été effectuée de telle sorte pour que de passage "mots de 32 bits → mots de 60 bits" soit aisé, et sans trop de perte dans l'efficacité.

FORTRAN IV *est un langage universellement répandu* : cela rend aisée une éventuelle adaptation du programme à des systèmes différents. Le temps de compilation est relativement court.

Le coût d'exploitation est faible ; la prévision étant de 200 secondes environ en Unité Centrale pour le traitement de 20.000 unités lexicales ("mots") données en entrée, dans le cas de la version dite "standard" (soit environ $ 60.00,/ soixante dollars).

Ce temps peut être réduit encore en transcrivant - ce qui est prévu ici, à l'Université de Colombie Britannique - certaines des parties du programme dans le langage Snobol (ou Spitbol). Nous voulons laisser le programme original en Fortran IV et laisser la possibilité aux différents centres qui utiliseront Texteconcord d'adapter eux-mêmes certaines parties du programme s'ils le désirent (et bien que cela ne soit pas nécessaire) à des langages différents tels que le Spitbol. L'utilisation de celui-ci permettra de réaliser un gain de 10% sur le coût prévu précédemment.

En résumé, il est prévu de livrer TEXTECONCORD sous les formes suivantes :

I. Entièrement écrit en FORTRAN IV-G. Le programme est directement exploitable sur tout système IBM ou CDC.

II. Ecrit partiellement en FORTRAN IV-G certaines parties requérant un temps de compilation relativement long étant transcrites en Spitbol. Du fait qu'il n'existe pas de compilateurs Spitbol aussi universaux que ceux du Fortran, le Centre devra – peut-être – adapter localement la version Spitbol.

*** Le programme TEXTECONCORD accepte tous les types de caractères (alphanumériques, alphabétiques, diacritiques, etc.) utilisés dans la translittération de données textuelles pour l'ordinateur.