

Au-delà du mot : la comparaison multilingue comme outil lexicographique

par

R.-F. POSWICK

Abbaye de Maredsous - BELGIQUE

755

Le "mot", notion intuitive ou scientifique ?

La dernière livraison des NOTES ON LINGUISTICS (N. 19, July 1981) publiées par l'Institut linguistique de la Wycliff Bible Society (Summer Institute of Linguistics, 7500 W. Camp Wisdom Road, Dallas, Texas 75236, U.S.A.) contient deux réflexions sur la notion de *mot*.

Velma Pickett qui travaille au Mexique sur la langue Zapotèque suggère que "tout type d'intuition impliqué dans le concept de *mot* est influencé par une tradition d'histoire du langage" (p. 33), elle continue "il n'existe pas de définition qui puisse être donnée du terme mot en général, incluant tous les facteurs grammaticaux, phonologiques et orthographiques traditionnels" (ibid.). Elle conclut en observant "qu'il se pourrait que la différence entre mot et phrase ne soit pas valable pour certaines langues" (p. 38). . . "mais ce serait cependant une erreur de ne pas garder le principe d'une distinction entre mot et phrase au plan pédagogique puisqu'on peut quand même déceler dans certaines langues et certains contextes, des différences psychologiques et structurales entre les deux" (ibid.).

Le second article de la même livraison, signé par Eunice V. Pike, se situe plutôt dans la perspective d'une analyse phonologique et rappelle les principaux moyens pour déterminer, selon le langage, ce qui peut être considéré comme un mot, "la phrase constituée par des syllabes pré- ou post- accentuées; l'accent; la durée du mot; les allophones ou allotones en dépendance de leur occurrence par rapport à l'environnement spécifique d'un mot; la distribution des phonèmes; l'harmonie des voyelles; les pauses possibles; et, enfin, la reconnaissance des morphèmes impliqués" (ibid., p. 39).

Cette analyse qui se réfère explicitement à une approche dans la tradition de Bloomfield en jouant beaucoup avec la notion de "free form" ou "forme libre", n'a pas la prétention d'apporter du nouveau, mais elle nous rappelle que la notion de mot n'est pas univoque ni parfaitement définissable.

Existe-t-il d'ailleurs une histoire de la notion de "mot" ?

Existerait-il, par exemple, une progression historique et méthodologique ? Elle partirait d'une approche lexicale héritée des grammaires classiques où le mot est surtout considéré comme un "son monosyllabique ou polysyllabique, composé de plusieurs articulations, qui a un sens" (Littré, 1885) ou comme le support d'une idée "Toute idée, simple ou complexe, se traduit par des sons, des groupes de sons, et des bruits, qui forment des mots, signes des idées (Brunot, *Pensée et Langue*, p. 3), définissant le mot comme "L'unité sémantique minima de la parole" (S. Ullmann, *Précis de Sémantique française*, p. 33). Cette progression s'orienterait ensuite vers une approche plus linguistique où le mot se définit plutôt "par le fait que le système d'articulations dont se compose le mot est susceptible d'être isolé ou déplacé dans la phrase sans cesser de répondre à un concept donné. . ." (Marouzeau, *Lexique de la terminologie linguistique*, p. 89).

Martinet rappelait cette difficulté à cerner le mot et préférait parler de *syntagme autonome* (*Eléments de linguistique générale*, p. 112-115), insistant sur le fait que "le caractère de syntagme autonome qu'une forme latine comme *hominī* partage avec ses équivalents modernes *for man*, *pour l'homme*,

para el hombre, est plus essentiel que son caractère de mot" (ibid., p. 115).

Parmi les critères retenus par les lexicographes, perpétuellement affrontés à cette question, on retient "la stabilité du rapport entre la suite formelle et la notion exprimée" . . . L'impossibilité "d'intercaler d'autres éléments dans la suite lexicalisée", la "fréquence d'apparition d'une suite donnée dans le stock lexical" ou encore "l'absence de liberté dans le choix des termes" (J.B. Marcellesi, *La Linguistique*, Encyclopédie Larousse, 1977, p. 195). Mais la tâche reste difficile : "si l'élément minimum porteur de signification, le 'morphème', . . . si la 'phrase', si le 'syntagme' sont convenablement définis, il n'en va pas de même pour ces unités qui figurent dans les dictionnaires, qu'on sépare dans l'écriture -mais pas toujours dans la parole- et qu'on appelle sans hésiter des 'mots' " (A. Rey, *Introduction au Dictionnaire des Expressions et Locutions*, Robert, 1979, p. V).

Ces différentes approches du mot, même si elles tiennent compte partiellement des apports de la linguistique comparée, se réfèrent plus généralement au fonctionnement interne à une langue donnée.

Et, dans une telle langue, on aboutit encore souvent à des critères qui pourraient sembler relever d'un arbitraire intuitif :

- "la tendance à ne pas séparer dans l'énoncé les monèmes qui *sont sentis* comme étroitement unis par le sens est trop naturelle pour qu'on n'en trouve pas des traces dans toutes les langues" (Martinet, *ibid.*, p. 114). (nous soulignons).
- "les règles de formation du discours (grammaire, au sens large) ne suffisent pas à produire, par exemple, *mettre cartes sur table*. Ni pour la forme (on "devrait" dire : *mettre les cartes sur la table*), ni pour le sens. Voilà un élément de la langue française qui fait partie du système même, du *code*; qu'il faut donc maîtriser en tant que tel . . ." (A. Rey, *op. cit.*, p. V).
- ou encore le titre même de l'article de V. Pickett cité : "*How intuitive is the 'word' ?*".

D'où l'idée d'utiliser des procédés de comparaison systématique sur des corpus importants pour donner une base expérimentale plus large et plus critique afin de cerner ces unités aux limites mal définies.

*

*

*

La comparaison automatisée des "mots".

On a déjà utilisé des processus de comparaison automatisée pour comparer les états d'un texte et cheminer vers l'histoire de la transmission d'un texte (stemma codicum) ou la préparation d'un texte critique (cf. Froger, *La critique des textes et son automatisation*, Dunod, 1967, et tous ceux qui ont amélioré cette ligne de recherche : P. Gilbert, P. Tombeur, etc . . .).

On a également utilisé la comparaison de textes pour déceler les influences littéraires d'un corpus à un autre (cf. J. Raben et D.V. Lieberman, "Text comparison. Principles and a program" dans *The Computer in Literary and Linguistic Studies*, 1976, pp. 297-308).

On a aussi essayé de pousser la recherche dans une ligne plus linguistique pour déterminer des états successifs de la langue (R. Tomassone et S. Monsonogo, "Déterminer un état de l'ancien français par la comparaison statistique de textes voisins ?" dans *Langue française*, 40, 1978, pp. 5-17.

Je ne connais pas, à ce jour, de tentative d'utiliser les possibilités de l'ordinateur pour comparer de larges corpus ayant des contenus identiques, en des langues différentes, pour y observer non seulement des mécanismes de "traduction", mais des "structures" lexicales.

*

*

*

Le travail de l'équipe de Maredsous.

Si le "mot" n'est qu'une abstraction arrêtée pour les facilités didactiques de la lecture, de la grammaire ou du lexique, l'humble travail qui consiste à suivre, sur un grand corpus de textes, l'ensemble complexe de la chaîne de signification et à la découper en fonction des plus petites unités réelles de sens représentées par les éléments du contexte dans les différentes versions du même texte, cet humble travail peut faire apparaître tout un mécanisme ou une typologie de la "chaîne significative" pour plusieurs langues différentes.

C'est à cela que nous aboutissons sans en avoir fait la théorie au préalable.

Notre premier but était de constituer les instruments informatiques qui nous permettraient de donner aux biblistes un nouvel outil, plus perfectionné, plus complexe, dans la tradition des grandes Concordances bibliques des siècles antérieurs.

Il s'agissait pour cela d'enregistrer tout le corpus biblique, dans les principales langues de sa transmission jusqu'à nous : hébreu/araméen, grec, latin . . . et français ou anglais -éventuellement syriaque et arabe.

Il s'agissait de créer la chaîne de traitement qui permettait de corriger, d'analyser, d'éditer chacun de ces textes de façon qu'ils soient parfaitement utilisables comme corpus indépendant selon les règles propres du lexique de chacune des langues.

(C'est ainsi que pourra paraître prochainement une Concordance de la Bible de Jérusalem, peut-être une autre en anglais en même temps qu'une bilingue en arabe et en grec, etc . . .).

Nous avons donc créé cette chaîne de traitement qui nous donne un ensemble cohérent de données dont les résultats peuvent être exploités sous les diverses formes souhaitées par le chercheur ou l'éditeur (y compris l'affichage sur écran en caractères spéciaux : grec, hébreu, arabe, syriaque, ou la sortie sur papier d'ordinateur dans les mêmes caractères, ou encore en typographie par composition électronique).

Algorithmes de comparaison.

Ce matériau énorme est en cours de comparaison systématique pour les 8 textes complets et les 5 langues que nous avons choisies pour être le noyau fondamental de cette Concordance Multilingue électronique.

- 1) Cette comparaison est fondée sur le postulat d'équivalence globale de contenu pour un segment de texte donné par les différentes versions dans ce corpus; -notamment, pour la Bible, le *verset* biblique ou une subdivision plus pertinente de celui-ci.
- 2) Elle est préparée par une lemmatisation interne ou absolue (1) que nous appelons "de premier niveau" et qui réduit les mots typographiques (séparés dans l'écriture par un blanc) à leur forme canonique telle qu'on peut la trouver dans un lexique ou un dictionnaire qui fasse autorité, de sorte qu'en réalité, l'unité analysée est parfois plus petite (*vobis-cum*; *can-not*) ou plus grande (à cause de; *s'en aller*) que le mot typographique. Certains de ces mots sont donc neutralisés puisqu'a priori, on peut penser qu'ils n'auront pas d'équivalent formel dans d'autres langues (comme l'absence d'article en latin).
- 3) L'équilibre des masses de mots ainsi rétabli, une série de programmes peuvent alors calculer et comparer les fréquences de chaque lemme séparément et la fréquence de ses co-occurrences avec ceux des autres versions du même texte; les correspondances les plus vraisemblables entre ces lemmes sont déterminées dans l'ordre décroissant de ces rapports de fréquences.
- 4) Ce processus qui, sans doute, n'est pas le meilleur ni le seul possible, réussit en moyenne à 75 %, parfois même 90 %, selon la proximité de langue, comme l'a montré une expérience entre le grec et l'arabe, ou selon la fidélité de traduction (le latin de la Vulgate, par exemple, est de loin le plus servilement respectueux de la matérialité de ses sources).

Une bonne part du résidu est provoquée par certains phénomènes repérés comme systématiquement "bruyants" : mots-outils trop fréquents ou à structure très particulière dans une langue donnée, voire même certaines "expressions de langue".

Avec ces derniers phénomènes, nous touchons aux frontières de notre propos : le problème du statut linguistique et de la classification des "mots-nonmots" qui déborde le cadre d'une lemmatisation purement lexicale.

L'expression (2) comme unité lexicale.

1. L'entreprise de comparer, lemme à lemme, 8 versions d'un même texte (la Bible) dans 6 langues différentes, chaque version étant lemmatisée au niveau du "mot typographique" (unité de texte inscrite entre deux blancs), amène d'emblée une constatation fondamentale : ce qui est exprimé par un lemme unique dans une version déterminée peut parfois être exprimé par la combinaison de plusieurs lemmes dans une autre version, qu'elle soit de langue identique ou différente.

Ex. : Mt 14, 19 : εὐλογέω = benedico = bless = bénir = rendre grâce = prononcer la bénédiction.

Cette constatation oblige alors à effectuer une lemmatisation de second niveau, c'est-à-dire à créer un lexique où l'unité de base (lemme) n'est plus nécessairement le "mot" typographique, mais peut être le groupe de mots ou "expression".

L'étude qui suit, basée sur le fichier comparé de l'évangile de Matthieu en grec (UBS 3), latin (Vulgate, éd. Weber), anglais (R.S.V.) et français : Maredsous (M), Jérusalem (J), Segond (S) et T.O.B. (T), voudrait suggérer comment le problème pratique de garder une cohérence maximum dans cette opération de relemmatisation rejoint en de nombreux points la problématique théorique de la lexicologie et de la linguistique contemporaines, et comment la comparaison multilingue systématique peut compléter utilement certaines méthodes comme la "méthode des commutations" ou l'analyse componentielle, et ainsi demeurer une méthode essentielle de la linguistique descriptive comme elle l'a été de la linguistique historique.

2. En premier lieu, la comparaison de lemme à lemme, en obligeant le compareur à dégager des formes composées le noyau lexical, fournit une méthode objective pour isoler et caractériser les *éléments flexionnels* (fonctionnant comme "outils grammaticaux" ou "morphèmes").

2.1. C'est ainsi que des éléments recensés par le lexique comme des lemmes au sens plein (le "lexème" de MARTINET), doivent être, à la lumière de la comparaison interlangue, considérés comme des "flexions externes", correspondant à des flexions, soit également externes, soit internes (le "morphème" au sens classique du terme) dans d'autres langues.

- Ex. : Mt 3, 11 : en **SIGNE** de pénitence (M)
 en **VUE** du repentir (J) } = εἰς (μετάνοιαν)
 pour **AMENER** à la repentance (S)
- Mt 4, 26 : de **PEUR** que (M, J, S)
 pour **EVITER** que (T) = μήποτε (ne forte, lest)

Il est frappant, à cette lumière, de remarquer combien la grammaire traditionnelle néglige la notion de "verbe auxiliaire", en la réduisant la plupart du temps, aux verbes "avoir" et "être", alors que la comparaison interlangue montre clairement que de nombreux autres verbes n'expriment souvent que des modalités (temps, aspect, mode, voix) du procès :

- Ex. : Mt 3, 12 διακαθαριεῖ τὴν ἀλώνα αὐτοῦ
 il VA nettoyer son aire (M, J, T)
- Mt 4, 2 ὑστερὸν ἐπειάσεν
 il FINIT PAR avoir faim (T)
- Mt 4, 25 ἠκολούθησαν αὐτῷ
 ils SE MIRENT A le suivre (J)
- Mt 6, 2 ὅπως δοξασθῶσιν ὑπὸ πάντων ἀνθρώπων
 that they MAY be praised by men
- Mt 9, 21 ἐὰν μόνον ἄψωμαι
 si j'ARRIVE seulement A toucher

2.2. A l'inverse, des éléments difficilement pondérables pourraient être détectés et recevoir un statut linguistique, une insertion précise dans le système grammatical, parce que correspondant à des éléments d'autres langues déjà recensés comme "outils". C'est le cas par exemple de certains procédés stylistiques mettant en jeu l'ordre des mots, ou l'alternance degré zéro/degré plein :

- Ex. : Mt 7, 10 ὑμεῖς πονηροὶ ὄντες
 TOUT mauvais QUE vous êtes (T)
- Mt 20, 30 ἀκούσαντες ὅτι Ἰησοῦς παράγει
 apprenant que C'ETAIT Jésus QUI passait (T).

C'est le cas aussi de certains signes de ponctuation. Il est frappant, par exemple, de voir combien souvent la conjonction grecque ὅτι est traduite par la ponctuation " : " dans des formules telles que :

- Mt 5, 22 ἐγὼ δὲ λέγω ὑμῖν ὅτι

eh bien, moi je vous dis :

cf. Mt 5, 28; 5, 32 . . .

3. Mais il reste que souvent le travail de comparaison ne peut réduire la forme composée en une combinaison "lemme + flexion externe", mais doit y reconnaître un noyau lexical constitué réellement par un amalgame de lemmes, formant une "expression" (ou syntagme figé) qu'il faut "chaîner" avec un lemme simple d'une autre version (3).

Une théorie linguistique de ces constructions a été proposée naguère par E. BENVENISTE (*Problèmes de linguistique générale* II, p. 171-176) qui propose de les appeler "synapsie". Le processus de comparaison procure, là aussi, un moyen objectif d'établir l'autonomie lexicologique de constructions que le linguiste a parfois peine à distinguer d'une construction syntaxique régulière (4).

La comparaison multilingue ne peut agir, bien sûr, comme un critère automatique et absolu, mais plutôt comme un indicateur et un moyen de contrôle du sentiment linguistique. Peut s'y ajouter un important élément statistique : il est clair qu'une "expression" reconstituée très fréquemment pour être chaînée à un lemme simple d'une autre langue, a beaucoup plus de chance de constituer réellement une unité lexicologique stable. Ceci montre encore l'intérêt pour le linguiste d'avoir à sa disposition de grands corpus multilingues comparés et automatisés.

3.1. C'est le cas par exemple des *faux passifs* comme :

ETRE ASSIS	(κάθημαι)	} sans correspondant actif.
ETRE POSSEDE	(δαιμονιζόμενος)	
ETRE PRIS DE FUREUR	(έθυμώθη)	

C'est le cas aussi de certaines "expressions" adjectivales :

Ex. : DE DEUX ANS	(διετής)
DE PEAU, DE CUIR	(δερμάτινος)
QUI NE S'ETEINT PAS	(άσβεστος)
SANS CULPABILITE	(ανάίτιος)
SIS AU BORD DU LAC	(παραθαλάσσιος)

ou substantives :

Ex. : FILLE AVISEE	(φρόνιμος)
FILLE INSENSEE	(μωρά)
EAU FRAICHE	(ψυχρός)

HOMME JUSTE } (δίκαιος)
 RIGHTEOUS MAN }

"expressions" que le lexique classique n'a pas l'habitude de retenir comme entités indépendantes.

De même, la comparaison impose l'autonomie lexicologique de :

<i>répondre de</i>	(ἔνοχος ἔστιν)	par rapport à <i>répondre</i>
<i>look for</i>	(προσδοκάω)	par rapport à <i>look</i>
<i>armé de</i>	(μετά)	
<i>fait de</i>	(ἀπό)	

3.2. Comme le fait remarquer BENVENISTE (*op. cit.*, p. 173-4) ce mode de production lexicologique, parallèlement à et de façon plus souple que les procédés plus classiques de dérivation et de composition, correspond à une démarche plus analytique (d'où sa productivité dans la nomenclature technique) se rapprochant assez bien d'une opération que la sémantique contemporaine appelle "analyse componentielle" (ou recherche des différents traits sémantiques formant l'unité lexicale).

Qu'est donc l'expression "chambre la plus retirée" (T) sinon une analyse componentielle *naturelle* de *ταμεῖον* (Mt 6, 6); de même, "place d'où partent les chemins" par rapport à *διέξοδος* (Mt 22, 9); "entonner un chant funèbre" par rapport à *θρηνέω* ou "bureau du paiement des taxes" par rapport à *τελώνιον* (Mt 9, 9).

Ainsi apparaît le fait que les traducteurs sont souvent amenés à faire, de façon impromptue et non systématique, une sorte d'"analyse componentielle naturelle", et la comparaison multilingue peut fournir, là aussi, une méthode objective, heuristique ou de contrôle, pour les méthodes systématiques d'analyse componentielle en linguistique.

3.3. Ces gradations, allant d'un ensemble structuré de traits ("bureau du paiement des taxes") à une unité lexicale simple *τελώνιον* pourraient peut-être fournir des modèles naturels à la sémantique générative, cherchant à démontrer les mécanismes de transformation de structures profondes sémantiques (constituées d'un ensemble de "traits sémantiques") en structures de surface (les phrases prononcées en fait), par la mise en oeuvre de règles syntaxiques et lexicales. Ceci d'autant plus que ces gradations apparaissent, non seulement dans la comparaison interlangue, mais aussi dans la comparaison "intertexte" dans une même langue.

Ex. : Mt 14, 6 G γενέσια
 L DIES NATALIS
 A BIRTHDAY
 T ANNIVERSAIRE
 J ANNIVERSAIRE DE NAISSANCE

S ANNIVERSAIRE DE LA NAISSANCE
M FETE ANNIVERSAIRE DE LA NAISSANCE.

4. Un dernier problème, de type plus proprement lexicographique, peut être abordé dans le cadre de cette réflexion : celui de l' "entrée" de dictionnaire réservée à ces expressions.

La comparaison multilingue peut imposer des entrées différentes, par exemple pour TEMPS (*χρόνος*), AU TEMPS DE (*ἐπί*), EN MEME TEMPS (*ἅμα*), EN MEME TEMPS, QUE (*μετά*), TEMPS CONVENABLE, TEMPS VOULU (*καιρός*).

Le problème reste de déterminer le "mot-vedette" qui servira d'entrée pour chacune de ces expressions. Intervient alors la notion de "poids sémantique" pour désigner tel ou tel composant à la fonction de "mot-vedette". Il est clair, en nous reportant à l'exemple ci-dessus, que les locutions AU TEMPS DE, EN MEME TEMPS . . . sont à trouver sous le substantif. Il n'en va pas de même pour TEMPS CONVENABLE, TEMPS VOULU, où l'adjectif intervient aussi de façon importante, dans la signification de l'ensemble, et où la solution lexicographique équitable, et donc scientifique, ne peut être que la *double entrée*, à TEMPS et à CONVENABLE.

Dans le même ordre d'idée, la comparaison multilingue inviterait à pousser plus à fond une innovation amorcée déjà dans des dictionnaires récents : la prise en compte des affixes (préfixes et suffixes) et des éléments de composés comme "entrée" de dictionnaire. Si "artisan de paix" demande logiquement une double entrée, pourquoi *εἰρηνοποιός* (Mt 5, 9) ne le demanderait-il pas ?

Comparer de même : GRAND PRETRE et *ἀρχιερεύς*
MENER VIE COMMUNE et *συνέρχομαι*
SITUE PRES DE LA MER et *παραθαλάσσιος*

5. Ces quelques exemples, partiels et trop sommairement analysés, veulent montrer comment la comparaison systématique de traductions multilingues pourrait apporter aux linguistes une méthode particulièrement intéressante d'accès aux phénomènes linguistiques.

D'autres exemples pourraient être développés, comme la détection automatique des synonymes, l'étude des phénomènes de redondance, la recherche des "universaux" linguistiques, la détermination de "champs" lexicologiques . . .

Dans cette perspective, on peut penser que le fichier biblique comparé, en chantier à Maredsous, qui comprendra plus de 5 millions d'occurrences lexicales réparties entre 8 versions en 6 langues différentes, fichier susceptible d'un enrichissement indéfini et prêt à être consulté informatiquement, offrira sans nul doute aux linguistes une mine de données et un instrument de travail précieux.

Que sera le Dictionnaire de demain ?

Les limites de nos observations sont trop évidentes pour qu'on ait à les souligner : il faut attendre d'en vérifier les hypothèses sur un stock d'expressions, locutions, syntagmes correspondant à une masse de texte plus large; il faudrait appliquer le même processus à des textes d'autre nature et augmenter le nombre et la variété des langues mises en comparaison.

Mais surtout, il faudra rendre accessible de façon utile et permanente les acquis progressifs qui seraient réalisés un peu partout dans une ligne similaire si elle s'avère constructive pour l'étude des unités du langage et pour l'amélioration des processus automatisés de traduction.

C'est sur une question que nous resterons donc : quel aspect devra avoir le Dictionnaire de demain ? Quelle part de celui-ci devra-t-elle être électronique, et quelle part publiée ? Sous quelles formes doit-on avoir accès à un matériau beaucoup plus complexe, au niveau des entrées lexicales, que celui des Dictionnaires traditionnels ?

A plusieurs de ces questions, les lexicographes ont déjà répondu. La pratique, de plus en plus répandue, des *Dictionnaires d'usage* en est témoin.

Pour notre part, nous voulons laisser la question ouverte : nous sortons tout doucement d'une tradition d'écriture et de lecture pour entrer dans un large réseau de représentations nouvelles, quand nous aurons suffisamment de matériaux, d'expériences bonnes et mauvaises sous les yeux, on pourra se former un jugement plus sûr.

NOTES

- 1) Par opposition à une lemmatisation externe ou relative (en fonction d'autres langues ou versions).
- 2) Vu la confusion terminologique en linguistique, l'équipe de Maredsous utilise un vocabulaire le plus neutre possible par rapport aux diverses théories en cours. C'est ainsi que nous appellerons "lemme" l'unité de base du lexique (l' "entrée" du dictionnaire) là où beaucoup, à la suite de MARTINET, parlent de "lexème"; nous parlerons de flexion (interne ou externe) là où l'on parle plus souvent de "morphème", et d' "expression" là où d'autres parleront de "syntagme figé", "lexie" ou "synapsie" (BENVENISTE).
- 3) S'agissant d'une traduction, il faut toujours, bien sûr, se demander s'il ne s'agit pas d'une *glose* du traducteur, comme en Mt 6, 19 où dans "les voleurs percent les murs" (T), il est clair que le traducteur rend $\delta\iota\omicron\rho\upsilon\sigma\sigma\omega$ par "percer" et ajoute en glose "les murs" pour une meilleure compréhension.
- 4) "Il est difficile de distinguer parmi les suites de mots, celles qui, dues au hasard, sont provisoires et celles qui sont assez stables pour être considérées comme des constituants de la phrase au même titre que les mots" (J.B. MARCELLESI, *Le lexique* in *La Linguistique*, Encycl. Larousse, 1977, p. 195).

TRAITEMENT DE TEXTE
 pour la
 Comparaison Multilingue Automatisée
 MAREDSOUS - 1981

