# UNE NORME DE DÉPOUILLEMENT TERMINOLOGIQUE EN LANGUE FRANÇAISE

## 1 DU DÉPOUILLEMENT TERMINOLOGIQUE

Confronté à un texte spécialisé, le terminologue souhaite découvrir les termes techniques qui désignent les notions contenues dans le domaine de connaissance qu'il a choisi d'étudier. Les moyens informatiques permettent aujourd'hui de réaliser un dépouillement systématique des mots utilisés dans un texte préalablement encodé. Le travail de fichage contextuel s'en trouve considérablement allégé, accéléré et, surtout, enrichi. Le chercheur ne sélectionne plus des contextes en fonction du hasard, de l'attention du moment, mais prend connaissance de tous les passages où le mot est utilisé. Le caractère rigoureux et exhaustif de l'informatique déteint sur son travail et l'amène à définir préalablement un certain nombre de concepts (mot, syntagme, trait d'union...) qui, auparavant, demeuraient dans un flou savamment entretenu.

Toute recherche terminologique assistée par ordinateur comporte au minimum deux étapes : la saisie et le dépouillement proprement dit. Ce document entend fournir une norme précise à cette activité et l'adapter au fonctionnement du logiciel *Micro Oxford Concordance Program* (*O.C.P.*)[1]. Cette norme a été définie dans le cadre de recherches terminologiques menées à l'Institut supérieur de traducteurs et interprètes. La première application a servi de matériau de base pour un exercice de description terminographique dans le cadre d'un cours de linguistique appliquée en seconde année. Les exemples cités sont extraits du manuel de navigation dépouillé à cette occasion. Il s'agit du troisième chapitre de la cinquième partie du *Cours des Glénans*, intitulé *La mer et ses mouvements* et consacré à l'approche hydrographique des courants marins[2].

## 2 LA SAISIE

Le texte que l'on souhaite analyser doit être disponible sur un fichier stocké selon la norme *ASCII*. Pour arriver à ce résultat, on pourra procéder à un encodage, à la récupération d'un fichier informatique préexistant sous un autre format, voire à une reconnaissance optique des caractères à l'aide d'un scanner. Le choix entre ces différentes solutions dépendra des circonstances, de la comparaison des coûts et méritera d'être discuté avec un informaticien.

Dès que le texte est saisi sous le standard *ASCII*, il peut être soumis au dépouillement. Toutefois, il ne répond pas encore aux normes et aux exigences d'un travail scientifique pertinent. Le linguiste ne peut se contenter de connaître la présence d'un mot dans un texte. Il convient de savoir ce qu'est un mot, où ce mot apparaît, comment l'auteur utilise ce mot, dans quel contexte, selon quelle fréquence, etc.

---

[1]. OXFORD UNIVERSITY COMPUTING SERVICE, *Micro Oxford concordance program*, Oxford, Oxford University Press, 1988, 220 pp.

[2]. *Le cours des Glénans*, Paris, Seuil, 1990, 1135 pp. Nous remercions le Centre nautique des Glénans et les éditions du Seuil qui ont considérablement facilité nos recherches en terminologie nautique en nous fournissant la version informatisée de la dernière édition de cet ouvrage de référence.

Le programme *O.C.P.* est capable de répondre à toutes ces questions à condition de munir le texte d'un apparat critique que la machine interprétera. Le texte saisi doit donc être codé en fonction des procédures d'interrogation que l'on souhaite mettre en oeuvre. La planche 1 illustre l'aspect que revêt un texte sous format *ASCII* lorsqu'il a été codé en vue de son traitement par le logiciel *O.C.P.*

## 2.1 Les clés

L'apparat critique est codé sous la forme de clés qui caractérisent le texte qui les suit et étendent leur signification jusqu'à l'apparition d'une nouvelle clé de la même catégorie. Les catégories retenues pour le dépouillement standard sont :

- type de caractère
- nature du texte
- nature du passage
- volume, tome, section...
- chapitre
- page

La clé doit être rigoureusement codée selon les principes attendus par le logiciel. Elles figurent donc entre les signes < > et comportent un code catégoriel (par exemple *C* pour *caractère*) suivi d'une variable normée (par exemple *I* pour *italiques*). Ainsi, un mot en caractères italiques isolé dans un texte en caractères romains sera précédé par la clé <C I> et suivi de la clé <C R>.

Exemple : "<C R> on nomme <C I> *storm surges* <C R> les ondes marines exceptionnelles"

Les clés ne sont pas comptabilisées comme mots dans les dénombrement statistiques opérés par le programme *O.C.P.* (*cf.* 3.2.1).

## 2.2 Référenciation et mise en page

### 2.2.1 PRINCIPE DE BASE : RESPECT DE L'ÉDITION

La saisie informatique vise la création de concordances et d'index, c.-à-d. d'outils de travail qui permettent un retour au texte original (*cf.* 3). Le chercheur doit pouvoir retrouver avec précision chaque mot rencontré. Le principe de base de la saisie étant de respecter la présentation de l'édition retenue, toutes les références chiffrées propres à l'oeuvre doivent être fournies à l'ordinateur : volume, tome, livre, chapitre, paragraphe, page et ligne. Nous décrirons ici une norme de dépouillement standard comprenant le volume, le chapitre et la page, mais le lecteur ne doit pas ignorer qu'il est possible de multiplier ces références. Chaque édition peut donc poser des difficultés heuristiques qui exigeront une adaptation de la norme standard.

### 2.2.2 LES LIGNES

Le texte *ASCII* respectera les retours à la ligne de l'édition originale. Ceux-ci ne doivent pas être codés d'une manière particulière, car le programme numérote automatiquement les lignes et "réinitialise" leur comptage dès que survient un saut de page. Les titres et sous-titres sont comptabilisés comme lignes, tandis que les lignes des illustrations, des encadrés et des notes font l'objet d'une numérotation indépendante (*cf.* 2.2.6 et 2.4).

D'un point de vue informatique, le mot est défini comme une suite de caractères incluse entre deux espaces blancs (*cf.* 2.5). Les mots qui sont coupés par un tiret en fin de ligne ou de page doivent donc être impérativement ressoudés afin d'éviter que l'ordinateur ne considère qu'il s'agit de deux mots différents. On prendra pour principe conventionnel que tout mot coupé doit être encodé en entier sur la ligne de sa dernière syllabe[3].

.....la couver-
ture bleue...

.....la
couverture bleue...

Cette règle est également d'application si l'une des formes d'un mot composé est rejetée au début de la ligne suivante.

.....cet attrape-
nigaud en a surpris plus d'un...

.....cet
attrape-nigaud en a surpris plus d'un...

Dans l'usage courant, on observe que certains mots peuvent s'écrire avec un trait d'union (*fourmi-lion*) ou en une seule forme (*fourmilion*). Dans le cas où de tels mots seraient scindés en fin de ligne, on les encodera avec un trait d'union sur la ligne suivante. Lors de l'établissement des concordances, on adoptera la forme la plus fréquente chez l'auteur concerné.

.....le fourmi-
lion est un insecte névroptère...

.....le
fourmi-lion est un insecte névroptère...

### 2.2.3 LA SUBDIVISION EN VOLUME, TOME, SECTION...

Toute subdivision supérieure au chapitre est introduite par la clé <V X> où V signifie le type (volume) et X la variable chiffrée (un caractère maximum). On "incrémente" cette clé lors de tout changement de volume, tome, section...

[3]. Plutôt que de ressouder le mot sur la ligne de sa première syllabe, on a choisi le rejet à la ligne, car il permet d'éviter de devoir encoder une ligne blanche lorsqu'une ligne ne contient que la deuxième partie d'un mot coupé.

#### 2.2.4 LA SUBDIVISION EN CHAPITRE

Cette subdivision est introduite par la clé <H X> où H signifie le type (chapitre) et X la variable chiffrée (deux caractères maximum). On modifie cette clé lors de tout changement de chapitre.

#### 2.2.5 LA PAGINATION

La pagination est introduite par la clé <P X> où P signifie le type (page) et X la variable chiffrée (quatre caractères maximum). Chaque nouvelle page suppose donc la présence d'une nouvelle clé <P X>. L'usage d'une clé de pagination indique au programme qu'il convient de réinitialiser le compteur des lignes.

#### 2.2.6 LA NUMÉROTATION DES LIGNES (ILLUSTRATIONS, CADRES ET NOTES)

Le programme numérote lui-même les lignes et recommence ce comptage dès qu'il est confronté à un nouveau type de texte (*cf.* 2.4) ou à un saut de page (*cf.* 2.2.2). A tout instant, le chercheur peut néanmoins modifier la numérotation : il lui suffit d'introduire la clé <L X> (L signifie "ligne" et X la variable chiffrée en deux caractères maximum) au début de la ligne concernée.

### 2.3 Typologie des caractères

Le terminologue est particulièrement attentif au point de vue qu'adopte le spécialiste par rapport aux mots qu'il utilise. La plupart des textes spécialisés respectent les principes scientifiques d'usage des italiques. La mise en valeur d'un terme peut également passer par l'usage des grasses, des soulignées et des capitales.

La clé "caractère" rend compte de ces pratiques pertinentes et s'introduit par la lettre C. Les codes suivants ont été retenus pour désigner les variables typographiques du *Cours des Glénans* :

- I = caractères italiques
- R = caractères romains
- G = caractères gras
- B = caractères bleus
- C = lettres capitales
- J = caractères gras et italiques
- K = caractères gras et capitales
- L = caractères bleus et italiques
- D = caractères bleus et capitales

Selon le principe de fonctionnement des clés, il convient de se rappeler que ces codes demeurent valables tant que l'ordinateur ne rencontre pas une nouvelle clé "caractère".

### 2.4 Type de texte

Un même ouvrage peut comporter des textes qui n'ont pas le même statut. Pour le terminologue, il est important de savoir si le mot est attesté dans le texte proprement dit, à l'intérieur d'un cadre, dans une note ou sous une illustration. La démarche heuristique le conduira à mélanger ces parties ou à les isoler dans des fichiers particuliers.

Quelle que soit la solution retenue, la clé "texte" rend compte de ces pratiques pertinentes et s'introduit par la lettre T. Les codes suivants ont été retenus pour désigner les variables :

- T = texte proprement dit
- C = cadre
- I = illustration
- N = note

### 2.5 Nature du Passage

La clé "nature" s'introduit par la lettre N et permet essentiellement de distinguer les titres et le texte écrit (de n'importe quel type).

- E = texte écrit proprement dit
- T = titre ou sous-titre

### 2.5 La ponctuation et l'apostrophe

Les signes de ponctuation sont reconnus par le programme et doivent donc être encodés normalement. Seuls le trait d'union et l'apostrophe exigent un traitement particulier.

On a pris pour principe de respecter systématiquement les traits d'union et les apostrophes utilisés par l'auteur. Cette attitude demeure valable même lorsque celui-ci semble s'éloigner de l'usage courant ou qu'il ne reste pas fidèle à ses propres habitudes.

*Papier-peint* = une seule forme
*Papier peint* = deux formes[4]

Un programme sera prochainement chargé d'analyser les formes unies par un trait d'union ou une apostrophe afin de décider si elles doivent être ou non séparées. Convenons de nommer *mot-ordinateur* toute suite de caractères comprise entre deux espaces blancs avant l'application du programme de découpage (par exemple : *s'écria-t-elle*). Le *mot découpé* est la suite de caractères comprise entre deux blancs obtenue après l'application du programme de découpage (*s' écria- t-elle*). Le programme sera donc conçu pour passer automatiquement du mot-ordinateur au mot découpé. Il sera capable de distinguer les traits d'union à fonction syntaxique (*suivez-moi*) et ceux qui possèdent une fonction lexicale (*poisson-chat*). Dès lors, l'encodeur ne devra pas se soucier d'introduire un espace blanc après les apostrophes et les traits d'union à fonction syntaxique.

| | | |
|---|---|---|
| *Suivez-moi* | deviendra | *suivez- moi* |
| *Poisson-chat* | restera | *poisson-chat* |

[4]. Le logiciel *O.C.P.* pourra être programmé pour regrouper ces exemples sous une même entrée au sein de la concordance.

Un même principe a été adopté pour les apostrophes. Le programme de découpage sera capable d'insérer un espace blanc après les apostrophes qui ont une fonction syntaxique (*s'abaisser*). Grâce à la consultation d'une liste d'exceptions régulièrement mise à jour, il sera également à même d'identifier les apostrophes qui marquent une unité lexicale (**grand'place*).

| | | |
|---|---|---|
| *S'abaisser* | deviendra | *s' abaisser* |
| *Grand'place* | restera | *grand'place* |

Même les cas complexes seront traités par ce programme.

| | | |
|---|---|---|
| *Qu'en dira-t-on?* | deviendra | *qu' en dira- t- on?* |
| *Le qu'en-dira-t-on* | restera | *le qu'en-dira-t-on* |
| *Suivez-moi jeune homme!* | deviendra | *suivez- moi jeune homme!* |
| *Le suivez-moi-jeune-homme* | restera | *le suivez-moi-jeune-homme* |

## 3 LE DÉPOUILLEMENT PROPREMENT DIT

La recherche de termes dans un texte se fonde sur quatre grands outils de dépouillement : l'index des formes, l'index des lemmes, la concordance des formes et la concordance des lemmes. L'index est une simple liste énumérant les mots du texte suivis des références permettant de les retrouver. La concordance fournit toutes les attestations d'un mot dans un texte, c.-à-d. tous les contextes dans lesquels le mot est identifié; la concordance fournit également des références précises et des calculs statistiques (*cf.* 3.2.1).

La distinction entre forme et lemme entend différencier les formes fléchies (variation en genre et nombre, conjugaison, variantes contextuelles) et les formes canoniques dites *lemmes* (entrées de dictionnaire). Index et concordances traitent le plus fréquemment les formes. Le travail sur les lemmes suppose la délimitation d'un vocabulaire clos pour lequel on identifiera toutes les formes acquises par le lemme dans le texte. Le présent document ne développera pas les problèmes de lemmatisation, c.-à-d. de regroupement des formes fléchies sous leur forme canonique[5].

### 3.1 Index des formes

Le programme *O.C.P.* qui réalise les index des formes est baptisé *INDEXF.CTL*. Il regroupe par ordre alphabétique toutes les formes identiques attestées dans le texte, fournit leur fréquence et mentionne ensuite les références essentielles (page et ligne) de tous les passages concernés. La planche 2 montre un extrait d'index.

[5]. *Cf.* VAN CAMPENHOUDT (M.), *Principes généraux pour le traitement informatique des textes français*, rapport inédit, 1988.

## 3.2 Concordance des formes

### 3.2.1 CONCORDANCE STANDARD

Le logiciel *O.C.P.* permet de définir "à la carte" l'aspect de la concordance. Le programme *CONCFS.CTL* (*concordance formes standard*) adopte une présentation classique où la vedette est centrée et suivie de la fréquence d'attestation. Les attestations sont centrées dans un micro-contexte d'une ligne et classées selon l'ordre alphabétique en fonction du premier mot qui suit la forme. Les références sont fournies à gauche du *listing* dans des "tuyaux" ou colonnes prédéfinies. La planche 3 illustre ce mode de présentation.

Chaque tuyau contient un type de référence : dans l'ordre, on trouvera le type de texte, la nature du texte, le type de caractère, le volume, le chapitre, la page et la ligne. Le programme remplit chaque emplacement en fonction des clés qu'il rencontre dans le texte.

Les barres obliques observables dans les contextes indiquent les endroits où se situent les retours à la ligne dans le texte original.

Tout à la fin de la concordance, on trouvera les données de statistique lexicale les plus fréquentes : fréquences relatives, nombre de formes rencontrées (étendue du texte) et nombre de formes différentes (vocabulaire), rapport entre ces deux données, etc.

### 3.2.2 CONCORDANCE INVERSE

Le terminologue sait que les termes techniques sont souvent des syntagmes. Ceux-ci transparaissent parfois mieux à travers un classement alphabétique des attestations en contexte, qui se fonde sur le mot précédant la forme concernée. La concordance inverse (planche 4) constitue donc un complément utile que l'on réalisera grâce au programme *CONCFI.CTL* (*concordance formes inverse*).

### 3.2.3 CONCORDANCE DES COLLOCATIONS

#### *3.2.3.1 SYNTAGMES À X MEMBRES*

Un terme technique composé est un syntagme figé (ou en voie de figement) composé d'au moins deux formes. Les programmes *CONCOL.CTL* réalisent des concordances qui fournissent les contextes où apparaissent des collocations de plusieurs formes. Le chiffre intégré dans le nom du programme indique le nombre de formes incluses dans le syntagme, soit deux (*CONCOL2.CTL*), trois (*CONCOL3.CTL*) ou quatre (*CONCOL4.CTL*). Le classement alphabétique de ces syntagmes s'opère, comme dans les concordances standards, en fonction du premier mot qui suit. La planche 5 fournit un exemple de concordance des collocations de quatre mots[6].

---

[6]. Les concordances de collocations ont évidemment une taille beaucoup plus élevée que les simples concordances standards. Le chercheur est le plus souvent à la recherche de marques de figement qu'il identifie en fonction de la fréquence d'attestation. Il pourra donc choisir de jouer sur les paramètres (*cf.* 3.3) et exiger seulement les collocations attestées plus d'une, deux... n fois.

#### *3.2.3.2 SYNTAGMES COMPORTANT UNE PRÉPOSITION*

La préposition *de* intervient dans de nombreux termes techniques composés, soit sous sa forme pleine, soit sous sa forme élidée (*d'*), soit encore sous la forme d'un article défini contracté (*du*, *des*). On a donc prévu un programme *CONCOLD.CTL* qui réalise une concordance des collocations de trois termes comportant la lettre *d* à l'initiale du deuxième terme. La planche 6 illustre l'intérêt d'une telle démarche.

### 3.3 Variantes

Les programmes mentionnés réalisent strictement les fonctions décrites dans ce texte. A condition de s'initier au fonctionnement d'*O.C.P.*, il est possible de modifier aisément divers critères d'analyse, telles les références, les clés, la fréquence des formes à prendre en considération, etc.

Ainsi notre intérêt pour l'oral nous a-t-il conduit à prévoir, dès à présent, une adaptation de chaque programme pour l'analyse de corpus oraux. Il s'agirait dans notre esprit, d'ouvrir une perspective qui observerait les langages spécialisés dans leur usage quotidien.

Marc Van Campenhoudt,
octobre 1990,
(publié dans *Equivalences*, n° 21/1-2, 121-136)