TRAITEMENT DES INFORMATIONS TEXTUELLES

LE PROCESSUS DE LECTURE

- Historiquement, le discours oral a longtemps été associé à l'écrit. Les premiers écrits étaient destinés à être lus par des " professionnels "; ils étaient constitués d'une suite de mots sans blancs ni ponctuation pour les séparer.

- Ce n'est que suite à l'invention de Gutenberg que des progrès sensibles ont été observés en matière de structuration de textes. Les progrès de l'écriture ont permis d'améliorer le rythme initial de la " lecture orale "

- Nous relatons brièvement quelques expériences faites entre 1800 et 1900 au sujet du processus de lecture visuelle.
  - L'expérience d'un *typographe* a fait ressortir la nette différence entre l'utilisation des anciens et nouveaux caractères (Didot/Garamond)
  - L'expérience d'un *notaire* a prouvé qu'en fait un texte contient beaucoup d'informations redondantes; que, par exemple, les mots pourraient être devinés par l'enchaînement du mot précédent et du mot suivant ou que la presque totalité des lecteurs ne lit que la partie supérieure des lignes.
  - L'expérience d'un *ophtalmologue* a conduit aux conclusions suivantes: le mouvement horizontal des yeux pendant la lecture se fait par saccades (1/4 à 1/3 de seconde pour la fixation, 1/40 de seconde pour la transition); la vitesse de lecture est fonction du nombre de mots et non du nombre de lettres.

- Ultérieurement, la théorie de l'information a corroboré les expériences faites et notamment le fait que le codage long est plus efficace. Toutefois, l'esprit humain se différenciant d'un ordinateur, l'aspect prévisible d'un texte accélère le processus de lecture.

- Enfin, il faut encore distinguer entre *lecture intégrale* et *lecture partielle*. Une lecture partielle peut être obtenue par divers phénomènes:
  - l'*ecrêtage linguistique*;
  - le *repérage* de débuts de phrases, d'éléments dans une liste, etc.;
  - la *recherche de points d'ancrage* (qui peuvent être concrétisés par des variations typographiques: utilisation de capitales, italique, gras, etc.)

COMPOSITION ET MISE EN PAGE

- L'origine des caractères est très lointaine, elle remonterait à plus de trois mille ans. Avant d'être gravées, elles étaient dessinées au pinceau et ceci est à l'origine de certaines propriétés encore apparentes de nos jours dans les polices de caractères utilisées (empattement, largeurs variables, force de trait...).

  Les capitales romaines ont ensuite subi de nombreuses modifications pour aller dans le sens d'une écriture plus petite. Les lettres furent reliées entre elles et l'écriture Caroline, datant de la fin du premier millénaire, est très voisine de nos minuscules. S'ensuivirent d'autres modifica-

tions telles que combinaisons des majuscules et minuscules, apparition des chiffres arabes et de la ponctuation.

- L'événement historique à l'origine d'un bouleversement, dans le domaine de la typographie, est l'invention, par Gutenberg, de la notion de caractère mobile qui a donné naissance à "l'écriture mécanique". Les caractères étaient fondus dans un moule commun dont le fond était tapissé de matrices en cuivre (caractère en creux), elles-mêmes frappées par des poinçons d'acier (caractère en relief). Le moule était réglable de façon à recevoir des matrices de largeur et de hauteur (corps) différentes, le texte était ensuite composé manuellement à partir de ces caractères mobiles. La composition d'un texte doit tenir compte de nombreux critères parmi lesquels:
  - la *largeur* du caractère métallique, appelée la chasse;
  - le *corps* du caractère qui détermine l'interligne (celui-ci doit être suffisant pour tenir compte des majuscules ainsi que des hampes des minuscules);
  - l'*approche*, c'est-à-dire, la distance horizontale qui sépare deux caractères.

  Les caractères métalliques sont tous inscrits dans le rectangle que constitue leur socle. Ceci représente un handicap si l'on veut combiner certaines lettres de manière esthétique comme cela était possible avec les caractères dessinés. Un exemple tout à fait caractéristique est la combinaison du A et du V. Ces deux lettres, côte à côte, provoquent un espace trop important. L'opération qui consiste à réduire l'espacement entre certains caractères s'appelle le crénage. Techniquement, le typographe pouvait procéder de deux façons: mortaiser les caractères ou fondre des combinaisons de lettres comme: "AV", "Ty", "ffl". Celles-ci sont appelées ligatures.

- La composition d'un texte nécessite le calcul de l'espace optimal entre les mots; ceci est important pour garantir une lecture confortable du texte (il s'agit, notamment d'éviter le phénomène de lézardes). L'interlettrage s'utilise essentiellement dans les lignes de capitales afin d'harmoniser leur approche particulièrement irrégulière. Enfin, l'interlignage, c'est-à-dire l'espacement entre les lignes a également son importance.

FORMATAGE DES DOCUMENTS

- La production d'un texte satisfaisant des exigences typographiques élevées nécessite le développement d'algorithmes relativement complexes. Il s'agit notamment de déterminer judicieusement les coupures de lignes dans les paragraphes, les sauts de page pour l'impression ou encore, le placement des figures flottantes.

  D.E Knuth et M.F Plass ont proposé un modèle et un algorithme qui satisfont bon nombre des critères; c'est notamment cet algorithme qui est utilisé dans le logiciel Latex.

Le modèle de Knuth et Plass

- Selon le modèle un texte est constitué d'une suite de mots (éventuellement coupés), d'espaces et de contraintes appelés respectivement:
  - *boîte* (box) dont le contenu est " inaccessible ";
  - *colle* (glue) qui permet de " coller " les boîtes entre elles en laissant un espace plus ou moins important;
  - *pénalité* (penalty) qui permet de déterminer la pénalisation due à la contrainte rencontrée.

- Une boîte est caractérisée par sa largeur (w)

La colle est définie par une largeur (w) et deux facteurs: l'extensibilité (x) et la compressibilité (z)

Une pénalité est caractérisée par une largeur (w) et une valeur représentant la valeur de la pénalité (p).

- En résumé, un paragraphe peut être représenté par une suite d'éléments x1, ..., Xn où chaque élément est caractérisé par un quadruplet (wi, yi, zi, pi) avec les propriétés suivantes:

  yi = zi = pi = o si x1 est une boîte

  pi = o si x1 est de la colle

  yi = zi = 0 si x1 est une pénalité

- Un point de coupure de ligne n'est autorisé que si x1 est de la colle et x1-1 une boîte ou x1 est une pénalité

- Exemple

  Soit le texte suivant

  La différence entre les Suisses et les oiseaux est que les oiseaux font leur nid alors que les Suisses nient leur fonds.

  Le même texte avec mention des coupures de mots possibles

  La dif~fé~ren~ce en~tre les Suis~ses et les oi~seaux est que les oi~seaux font leur nid alors que les Suis~ses ni~ent leur fonds.

  Le même texte schématisé

  boîte (La) colle boîte (dif) pénalité boîte (fé) pénalité boîte (ren) pénalité boîte (ce) colle boîte (en) pénalité boîte (tre) colle ...

  Enfin, le texte modélisé

  boîte (f(La)) colle (6,3,2) boîte (f(dif)) pénalité (5,50) boîte (f(fé)) pénalité (5,50) boîte (f(ren)) pénalité (5,50) boîte (f(ce)) colle (6,3,2) boîte (f(en)) pénalité (5,50) boîte (f(tre)) colle (6,3,2) ...

Algorithmes de formatage

- Un premier algorithme possible (First Fit) consiste à former une ligne après le premier mot permettant d'en créer une de bonne longueur. L'inconvénient de cette méthode est que le facteur de compression des espaces peut fortement varier d'une ligne à l'autre.

- Un second algorithme (Best Fit) consiste à envisager toutes les solutions admissibles pour une ligne et choisir celle dont le coefficient de compression, en valeur absolue, est proche de 0. L'inconvénient de cette méthode est que le paragraphe est formé ligne après ligne; ce qui ne permet pas de répondre à certaines exigences typographiques.

- L'algorithme de Knuth (Optimum Fit) consiste à chercher tous les emplacements admissibles pour une coupure de ligne dans le paragraphe et à leur attribuer un coefficient de non-satisfaction. Ce n'est qu'à la fin du paragraphe que le choix définitif sera opéré en tenant compte des facteurs de non-satisfaction.