



# Web Référence

## Référencement sur mesure de sites internet

<http://www.web-reference.fr>  
18, rue Jean-François Leca  
13002 Marseille

contact : Caroline Sarrion : [caroline@web-reference.fr](mailto:caroline@web-reference.fr) - 04 91 91 67 27

### >Dossier actualité>> 1 - Vous avez dit Langage Naturel ?

Si le Langage Naturel semble être porté par l'air du temps, il est frappant de constater à quel point l'expression elle-même " langage naturel " manque sacrément de naturel. Je me suis prise à relever, petite manie de linguiste oblige, par quelles périphrases l'internaute, au moment d'entamer une recherche, était invité à faire une requête sans trop se poser de questions sur la bonne façon de poser une question :

"langue naturelle", "langage courant", "langage ordinaire", "texte libre". Frédéric Plais (Infoclic) dans "Chasseurs de Moteurs" (n°25) : nous dit que "naturel "signifie dans ce contexte " sans contrainte "! Infoclic invite l'internaute à "poser des questions avec ses propres mots".

Car c'est bien le propre d'une interrogation en langage naturel que d'offrir la possibilité à l'internaute de ne pas avoir à construire une requête en manipulant les opérateurs de la syntaxe booléenne ni à choisir les bons mots clé.

Les **interfaces de dialogue en langage naturel** qui font une entrée en force sur le marché français, invitent sans complexe l'internaute à s'exprimer en langage naturel. Et là, paradoxalement, dès que le dialogue s'engage avec une hôtesse robotisée, la conversation devient naturelle. En d'autres termes, plus l'illusion de la convivialité est poussée, plus le naturel du langage revient au galop (les bases de connaissances linguistiques de ces outils prévoient d'ailleurs toujours un bon dictionnaire d'insultes d'ailleurs!).

L'invitation de **Question.fr** à poser ses questions en langage naturel contribue à semer la confusion : ces questions sont recueillies et lues par des experts humains (c'est plus le principe d'une boîte aux lettres que d'un formulaire) et donc traitées humainement et non par des automates.

La différence entre les outils d'interrogation se réclamant du langage naturel et les autres se fait du côté des dispositifs mis en place pour interpréter les questions. Rien de vous empêche de poser une question en langage naturel avec des outils de recherche traditionnels (annuaires ou moteurs de recherche). Mais attention, tout ce que vous tapez ne sera jamais identifié que comme des **chaînes de caractères**. Ces outils affichent même parfois " supporter le langage naturel ". L'ambiguïté de l'expression est fort jolie : si l'on prend " supporter " dans son acception technique un peu anglicisante, on peut entendre là qu'il est possible de poser une question en langage naturel sans planter son ordinateur. Si on la prend dans son acception courante, on supporte quand on n'aime pas, n'est-ce pas ?. (A quand le fan-club du langage naturel ?) Plus sérieusement, il semble que c'est la formule consacrée des outils qui ne sont pas spécifiquement prévus pour traiter des requêtes en langage naturel mais qui n'opposeront pas de résistance farouche à mouliner une requête de plus de 2 mots.

Ces chaînes de caractères sont recherchées dans la base de données des outils, après :

- élimination éventuelle des **mots vides**
- avec repérage des **opérateurs booléens** éventuels
- par défaut, ajout d'un des opérateurs ET ou OU.

Finalement, l'interprétation d'une requête est du " Traitement automatique du langage naturel " lorsque ces chaînes de caractères sont identifiées comme étant des **mots**. C'est à dire des objets porteurs de sens.

Liliane Khouri  
Responsable du développement linguistique

Autre article disponible sur le site : 2/ Une interface de dialogue : Semantia  
<http://web-reference.fr/actu/semantia.html>

## >Glossaire

### Interfaces de dialogue en langage naturel

Agent conversationnel (ce terme est proposé dans le lexique d'AgentLand); en anglais, chatterbot. Ce type d'interface propose une simulation de dialogue par l'intermédiaire d'un personnage virtuel, animé ou non. Exemples : Julie (nomino), Sémantia ([www.web-reference.fr/actu/semantia.html](http://www.web-reference.fr/actu/semantia.html)), Cybelle (Cybion). Pour en savoir plus : le site [www.agentland.fr](http://www.agentland.fr) dédié aux agents intelligents.

### Chaîne de caractères

Séquence de lettres séparées par un blanc (un espace typographique). "olkdflezp" est une chaîne de caractères au même titre que "annuaire".

### Mots vides

Mots ayant peu de signification. La liste de ces mots constituent un "anti-dictionnaire". Ils sont également appelés "mots creux" (Voilà), "mots banals" (Noematics). Notons que justement dans le contexte du langage courant en français ils possèdent un poids très important qui peut changer radicalement le sens d'une phrase. La préposition "de" étant le "mot vide" le plus célèbre, point n'est besoin d'être linguiste pour voir la différence par exemple entre "pomme de terre" et "pomme par terre".

### Opérateurs booléens

L'adjectif " booléen " vient, rappelons-le tout de même, du nom du mathématicien Boole. Au delà de l'histoire, ces opérateurs basés sur la théorie des ensembles ET = intersection , OU = union , ont été utilisés au sein de la logique des propositions. Cette logique permettant de faire des démonstrations formelles pas seulement mathématiques mais aussi philosophiques ou tout simplement de formaliser des petits raisonnements comme :

A = le chat est dehors B = il fait nuit

A et B = Le chat est dehors et il fait nuit

A et B est vraie si A est vraie et B est vraie

Si A et B est vraie alors on peut en déduire que A est vraie et que B est vraie.

Ils ont également bien servi lorsqu'il a fallu sophistiquer les requêtes des systèmes d'interrogation de base de données pour effectuer des recherches croisées. Si on les trouve aujourd'hui sur la plupart des outils de recherche grand public, c'est pour permettre aux internautes de saisir plus d'un mot. Il faut savoir que dès que vous saisissez plus d'un mot le moteur qui va chercher les chaînes de caractères y associe forcément un ET ou un OU par défaut. La logique voudrait que tous les mots (ET) soient au moins recherchés en priorité mais ce n'est pas toujours le cas. Le OU ( qui est un ou inclusif - " cherche moi l'un ou l'autre ou les deux à la fois " - engendre évidemment beaucoup de bruit. L'opérateur SAUF qui est la traduction du " non " booléen peut être utile à condition que l'on ait une idée a priori des réponses possibles ou des ambiguïtés de la question, afin d'éliminer ce qui ne correspond pas à notre attente. Peu d'internautes l'utilise. Il est plus utile dans le contexte d'une deuxième passe d'élimination sur un " lot " de réponses, il fait alors fonction de filtre, ou de soustraction qui est proposé par exemple chez [www.Kartoo.fr](http://www.Kartoo.fr) sous la forme d'un signe négatif (-).

### Mot

Attention! Terrain glissant! Ce concept est l'objet d'innombrables définitions linguistiques. Rappelons simplement que l'identification des mots dans un dictionnaire permet de prendre en compte le fait qu'il soit simple ou composé ("expressions", ou "unités lexicales complexes") et de rendre aux mots vides une certaine dignité (on ne peut pas éliminer le "de" de pomme de terre sans créer de contresens).