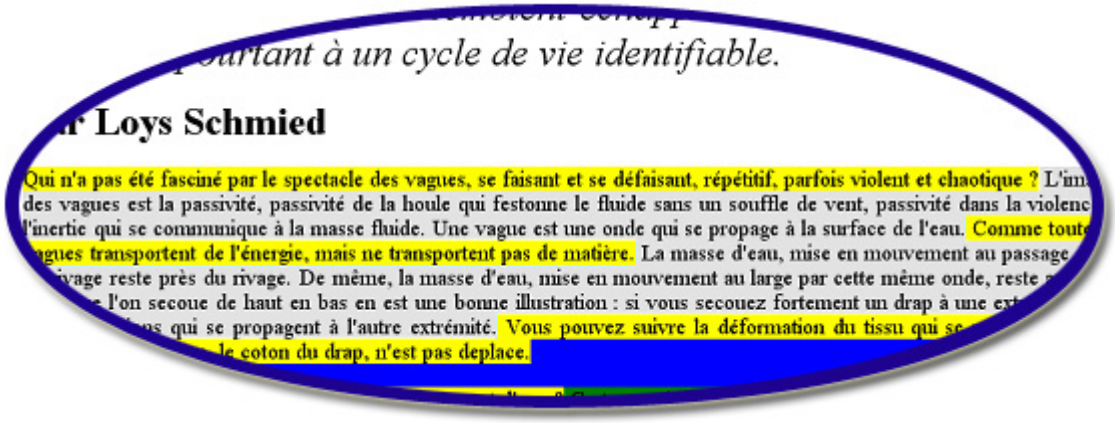
UFR de Sciences,
Section Informatique

Master RADI

Projet annuel

Détection d’explications à l’intérieur d’un document



Julien Van Den Bossche
contact@julienvdb.com

Sous la direction de Nadine Lucas
nadine@info.unicaen.fr

Sommaire

1 Introduction

Dans les diverses applications de recherche documentaire, les textes de type explicatif ont un intérêt particulier. Le but de ce projet est de voir comment les détecter. Ces textes explicatifs ont une structure particulière qui va nous permettre de détecter et délimiter les passages de textes explicatifs.

Une explication est un passage développant un sujet qui est amené dans un texte. L'explication est amenée par ce que l'on va nommer le « posé ». Ce posé précède généralement l'explication. Nous allons donc travailler sur le couple « posé / explication ».

Nous verrons que la détection des couples « posé / explication » se base sur des indices typographiques, grammaticaux et des règles positionnelles.

Nous travaillerons à différents niveaux, à différentes unités typographiques et nous verrons que l'explication peut être globale ou locale, relative à un petit segment de texte à l'intérieur du document. Nous verrons donc qu'il existe plusieurs niveaux d'explications dans un texte.

Dans ce rapport, nous allons donc vous présenter le sujet en détails pour ensuite vous montrer les différentes étapes du travail effectué.

Le code source, le rapport, sont disponibles ici :
http://www.julienvdb.com/universite/master/projet

2 Le contexte, présentation du travail à réaliser

2.1 Les contraintes, nos outils de travail

2.1.1 Les contraintes

Nous devons travailler avec du corpus au format HTML qui peut être en anglais ou en français.

Le code source HTML ne respecte pas obligatoirement les standards de la W3C (http://www.w3c.org) et l'on doit donc travailler avec du corpus dont le code HTML est mal formé.

En sortie, après avoir analysé notre corpus, nous devons donner le type de texte que l'on a pris en entrée :

- ✓ Soit il n'a pas de valeur explicative.
- ✓ Soit il est explicatif pour un sujet donné.

Nous allons voir ce que nous allons renvoyer à l'utilisateur dans ces deux cas.

2.1.2 Le texte non explicatif

Avant de détecter les explications dans un texte il faut d'abord connaître si le texte que l'on étudie fournit des explications par rapport à un sujet.

Ci-dessous un exemple de texte qui n'a pas de valeur explicative, c'est un conte de Grimm.

C'est une histoire mais ce n'est pas un texte qui s'interroge sur un sujet donné.

Un texte de ce type sera écarté par notre algorithme ou plutôt sera classé comme texte n'ayant pas de valeur explicative.

Les lutins

Une mère avait eu son enfant enlevé du berceau par les lutins, qui avaient mis à sa place un petit monstre à grosse tête avec le regard fixe, occupé seulement de boire et de manger. Dans sa détresse, elle alla demander conseil à sa voisine, qui lui dit de porter le petit monstre à la cuisine, de l'installer devant la cheminée et d'allumer le feu pour faire bouillir de l'eau dans deux coquilles d'œuf : " le monstre ne pourra pas s'empêcher de rire, lui dit-elle, et dès l'instant qu'il rit, c'en est fini de lui."

La femme fit tout ce que sa voisine lui avait dit de faire, et Grosse Tête, en la voyant mettre l'eau à bouillir dans des coquilles d'œuf, parla :

Moi qui suis vieux pourtant
Comme les bois de Prusse
Je n'avais jamais vu cuisiner dans un œuf!

Et le voilà qui éclate de rire, et il riait encore quand déjà surgissait toute une foule de lutins qui rapportèrent le véritable enfant, l'installèrent devant le feu et emportèrent avec eux le monstre à grosse tête.

Fig.1 Exemple d'un texte sans valeur explicative

Dans ce cas, nous informerons l'utilisateur que le texte n'a pas de valeur explicative et notre recherche s'arrêtera à ce niveau.

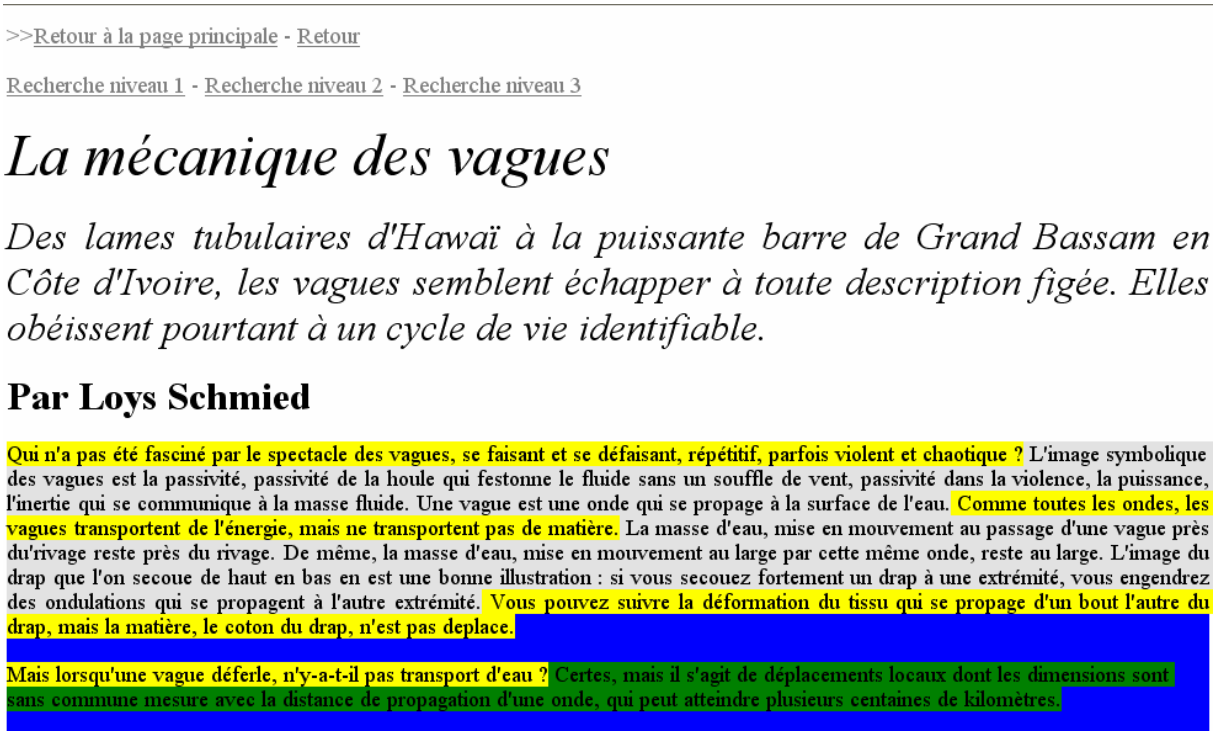
2.1.3 Le texte à valeur explicative

Dans ce cas nous devons fournir à l'utilisateur un texte colorié. Le coloriage représente les explications à l'intérieur du texte.

Nous verrons, plus bas dans ce rapport, qu'il y a plusieurs niveaux d'explications dans un texte selon l'unité typographique que l'on choisit.

L'utilisateur aura donc la possibilité de naviguer sur la page de HTML de sortie, par clic sur des liens pour faire apparaître ou disparaître les différentes explications marquées selon leurs niveaux.

Exemple :



>>Retour à la page principale - Retour

Recherche niveau 1 - Recherche niveau 2 - Recherche niveau 3

La mécanique des vagues

Des lames tubulaires d'Hawaï à la puissante barre de Grand Bassam en Côte d'Ivoire, les vagues semblent échapper à toute description figée. Elles obéissent pourtant à un cycle de vie identifiable.

Par Loys Schmied

Qui n'a pas été fasciné par le spectacle des vagues, se faisant et se défaisant, répétitif, parfois violent et chaotique ? L'image symbolique des vagues est la passivité, passivité de la houle qui festonne le fluide sans un souffle de vent, passivité dans la violence, la puissance, l'inertie qui se communique à la masse fluide. Une vague est une onde qui se propage à la surface de l'eau. Comme toutes les ondes, les vagues transportent de l'énergie, mais ne transportent pas de matière. La masse d'eau, mise en mouvement au passage d'une vague près du rivage reste près du rivage. De même, la masse d'eau, mise en mouvement au large par cette même onde, reste au large. L'image du drap que l'on secoue de haut en bas en est une bonne illustration : si vous secouez fortement un drap à une extrémité, vous engendrez des ondulations qui se propagent à l'autre extrémité. Vous pouvez suivre la déformation du tissu qui se propage d'un bout l'autre du drap, mais la matière, le coton du drap, n'est pas deplace.

Mais lorsqu'une vague déferle, n'y-a-t-il pas transport d'eau ? Certes, mais il s'agit de déplacements locaux dont les dimensions sont sans commune mesure avec la distance de propagation d'une onde, qui peut atteindre plusieurs centaines de kilomètres.

Fig.2. Le texte de sortie, avec les parties explicatives détectées.

2.1.4 Les outils utilisés

Pour réaliser ce travail nous utilisons les technologies HTML / CSS / PHP / MySQL.

CSS permet de réaliser des feuilles de style qui vont nous permettre de colorier notre texte de manière assez simple. Avec ces feuilles de styles nous distinguerons bien le fond de la forme.

MySQL va nous permettre de stocker l'information à chaque étape de notre analyse. En effet, à chaque étape, après chaque segmentation, donc après chaque modification sur le texte nous stockons la nouvelle information dans des tables de notre base de données.

PHP, qui fonctionne avec un serveur Apache a été choisi car il est assez simple d'utilisation, s'interconnecte bien avec une base de données MySql et possède un module de gestion des expressions régulières qui est assez complet.

En effet les expressions régulières vont être des outils indispensables pour la reconnaissance de modèles typographiques, syntaxiques…

2.2 Présentation de l'interface

Nous allons vous présenter l'interface qui va interagir avec l'utilisateur qui souhaite analyser un corpus. Ci-dessous sont présentées les différentes étapes de l'analyse d'un texte.

L'utilisateur doit choisir un fichier HTML. Il le fait en cliquant sur le bouton « parcourir »
Le programme vérifie l'extension du fichier et indique à l'utilisateur de re-choisir un fichier si ce dernier n'était pas un fichier au format HTML.

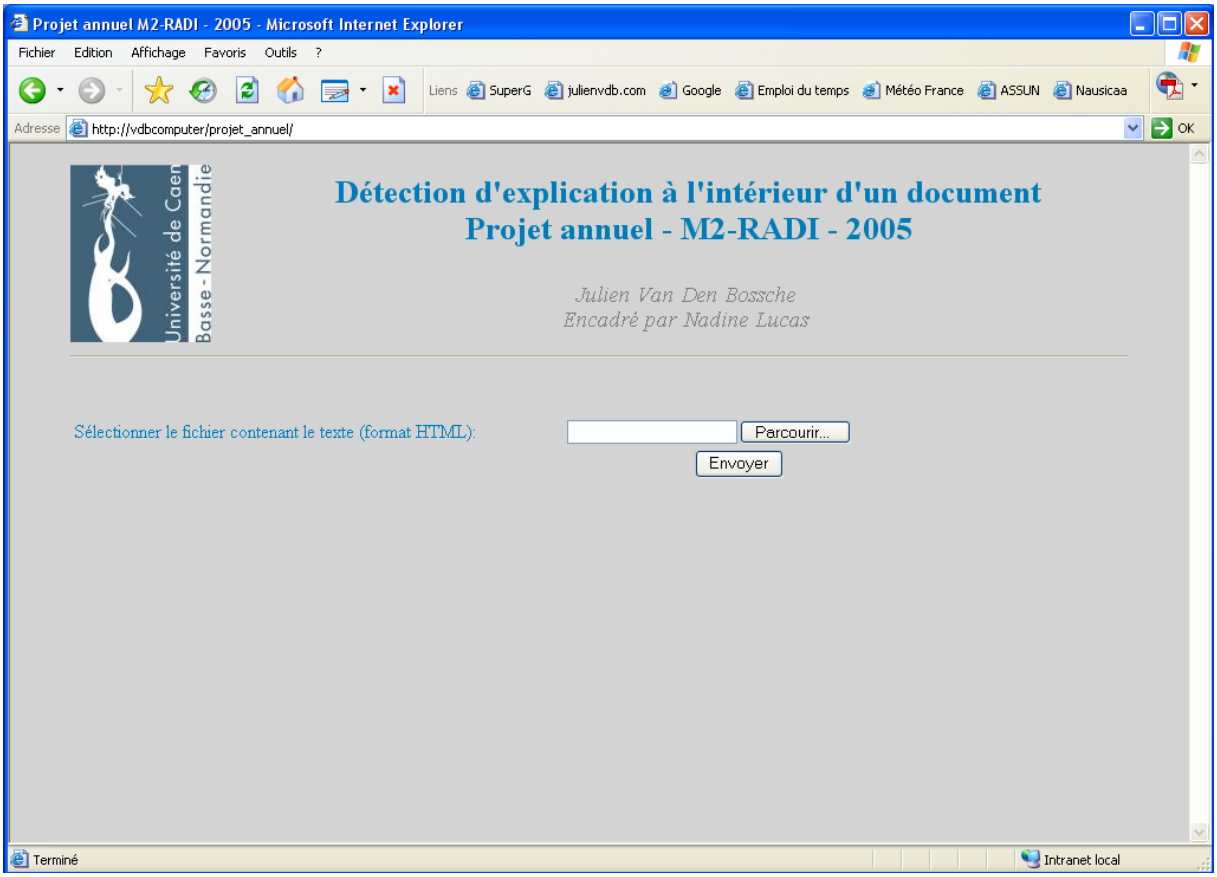


Fig.1. Choix du fichier source

Une fois le fichier téléchargé, l'utilisateur peut voir son fichier avec les différentes parties trouvées par le programme.
L'utilisateur peut alors choisir de supprimer des parties pour la suite de son analyse.
Il doit cocher les cases correspondant aux parties qu'il ne souhaite pas garder. Une fois le choix effectué, il doit cliquer sur le bouton en bas de la page.

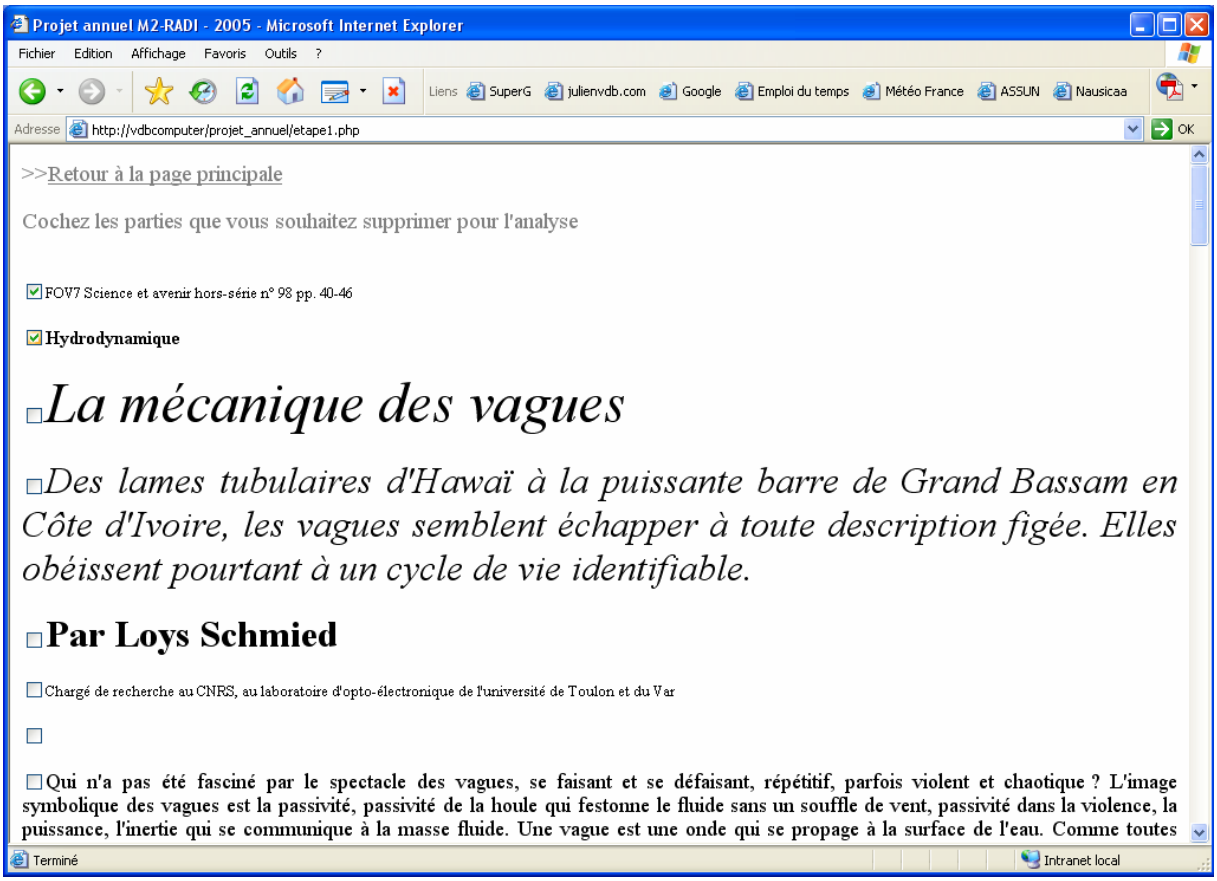


Fig.2. Choix des parties pour l'analyse

Une fois les parties sélectionnées, l'utilisateur est dirigé vers une page permettant de voir la segmentation du texte à différents niveaux.

Dans cette page l'utilisateur a la possibilité de repérer le corps de texte.

Un lien permet de passer à l'étape suivante qui permettra de détecter les marqueurs dans le texte.

Si le texte n'a pas pu être reconnu par le parseur HTML alors l'utilisateur sera prévenu que la structure se son texte n'a pas pu être acceptée l'outil d'analyse.

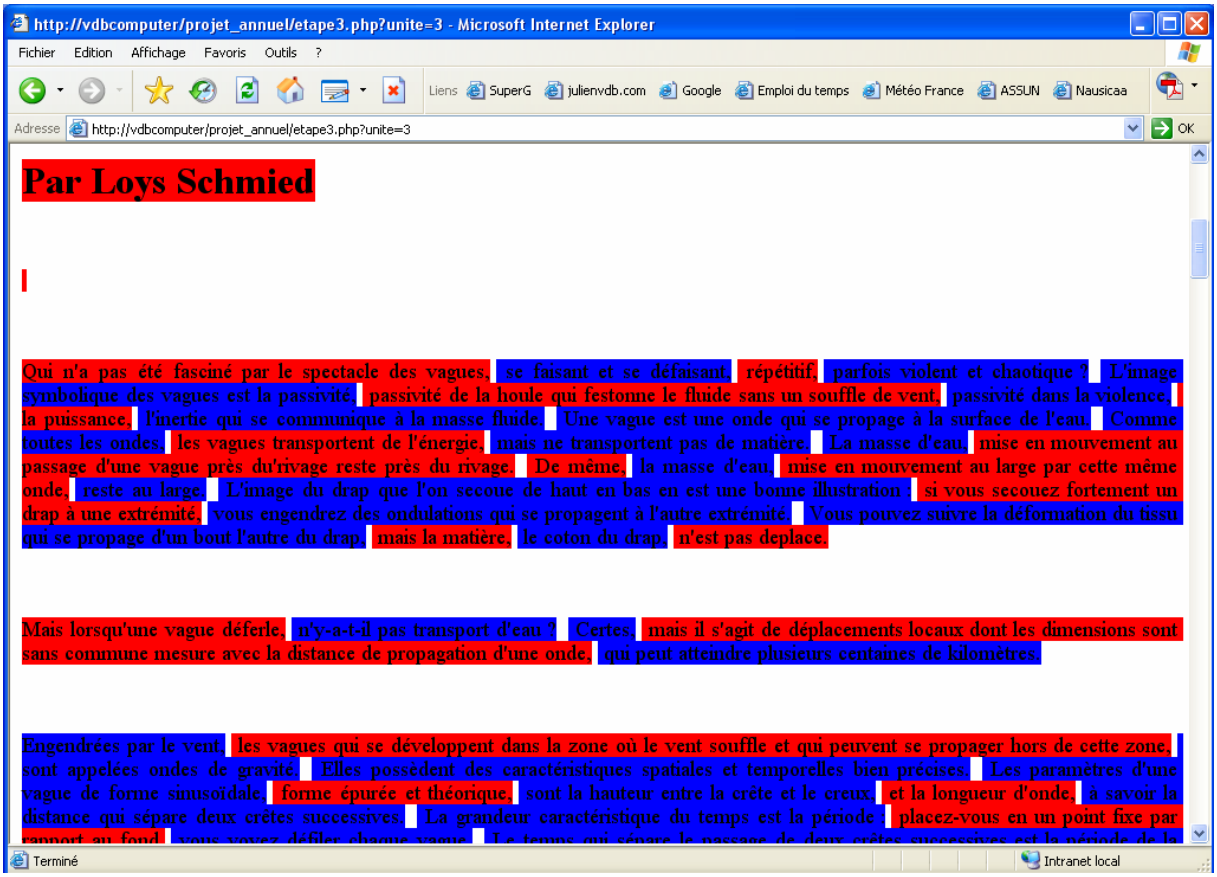


Fig.3.Etape de segmentation : ici le niveau est le virgulot

La dernière étape de l’analyse du texte est l’étape de détection des marques et de mises en relations de ces dernières pour ainsi détecter les explications dans le texte à différentes unités typographiques.

La page d’accueil de cette étape nous montre les éventuelles explications de niveau 1 (unité de base du texte).

Si le texte n’est pas explicatif, l’utilisateur sera prévenu.

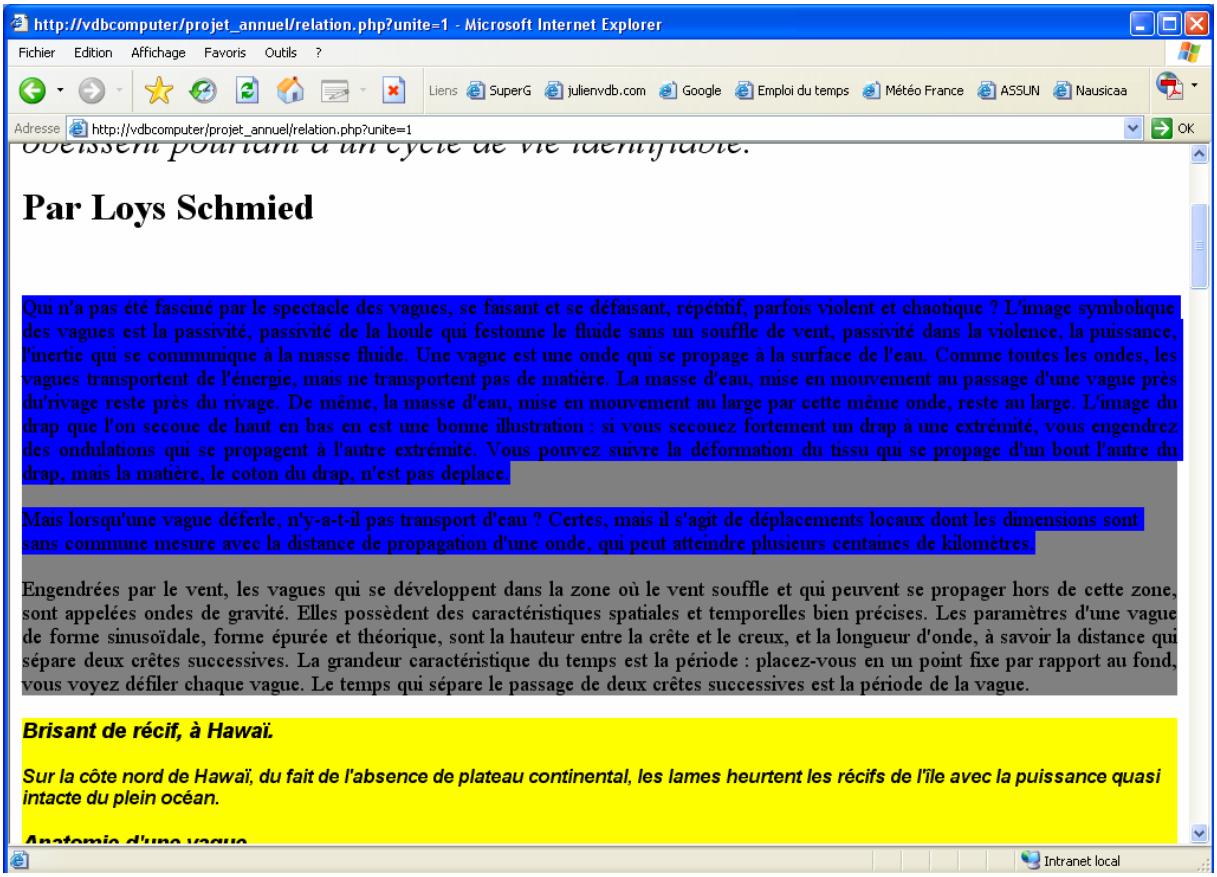


Fig.4. Etape de visualisation de la segmentation : ici segmentation de niveau 1 avec le posé en gris et en dessous, en jaune, l'explication pour ce posé.

Si l'utilisateur clique sur une unité typographique il verra en détails les explications locales pour cette partie.

Le détail pour cette partie est l'analyse des relations entre marques avec une unité typographique plus petite.

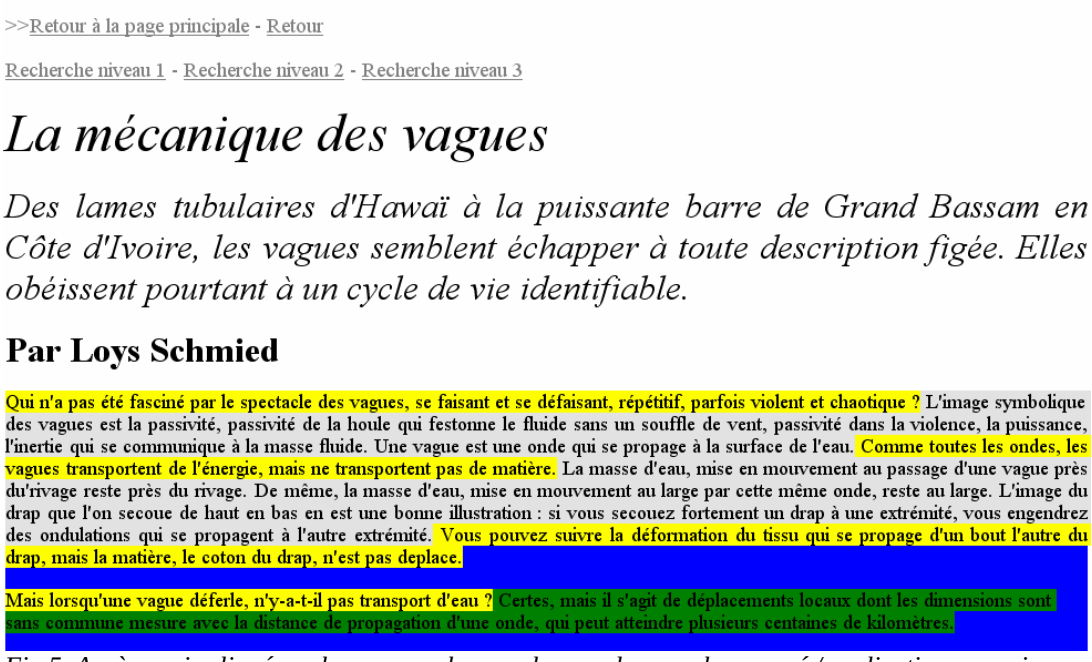


Fig.5. Après avoir cliqué sur le paragraphe, on observe des couples « posé / explication » au niveau de la phrase.

Après cette première partie de présentation du travail à réaliser nous allons vous expliquer quels moyens nous avons utilisés pour parvenir à nos fins. Dans la suite de ce rapport seront donc présentées les différentes étapes de réalisation du travail.

3 Première étape, étude de corpus

Pour me familiariser avec la notion de texte à valeur explicative il m'a fallu regarder beaucoup de corpus pour reconnaître ainsi « humainement » un texte explicatif d'un autre.

Dans cette étape je me suis aperçu que certains indices typographiques étaient importants pour la détection d'explications mais j'ai remarqué que l'organisation générale du texte, sa mise en forme matérielle était importante.

Il me faut détecter l'unité typographique de base du texte pour ainsi découper le texte avec une unité typographique plus petite et ce jusqu'à l'unité typographique virgulot.

S'occuper de la mise en forme matérielle signifie donc regarder le code source HTML et ses balises.

J'ai donc effectué cette tâche sur chaque corpus pour regarder comment le HTML était formé.

Sans trop de surprises, je me suis aperçu que ce dernier ne respectait aucunement les standards de la W3C (http://www.w3c.org) et que le balisage ouvert n'était pas forcément fermé ou n'était pas fermé au bon endroit.

La première étape de mon travail a alors était de réfléchir comment travailler au mieux dans un code source mal formé.

4 Travail sur le code source HTML.

4.1 Pourquoi un travail sur la mise en forme matérielle ?

Le but étant de travailler avec le corpus que l'on possède, il ne s'agit pas de modifier le code source HTML pour essayer de l'améliorer.

La partie qui nous intéresse dans le code source sont les éléments compris entre les balises <BODY> et </BODY>.

Si la balise <BODY> n'existe pas nous pouvons chercher la balise de fermeture </HEADER> si elle existe.

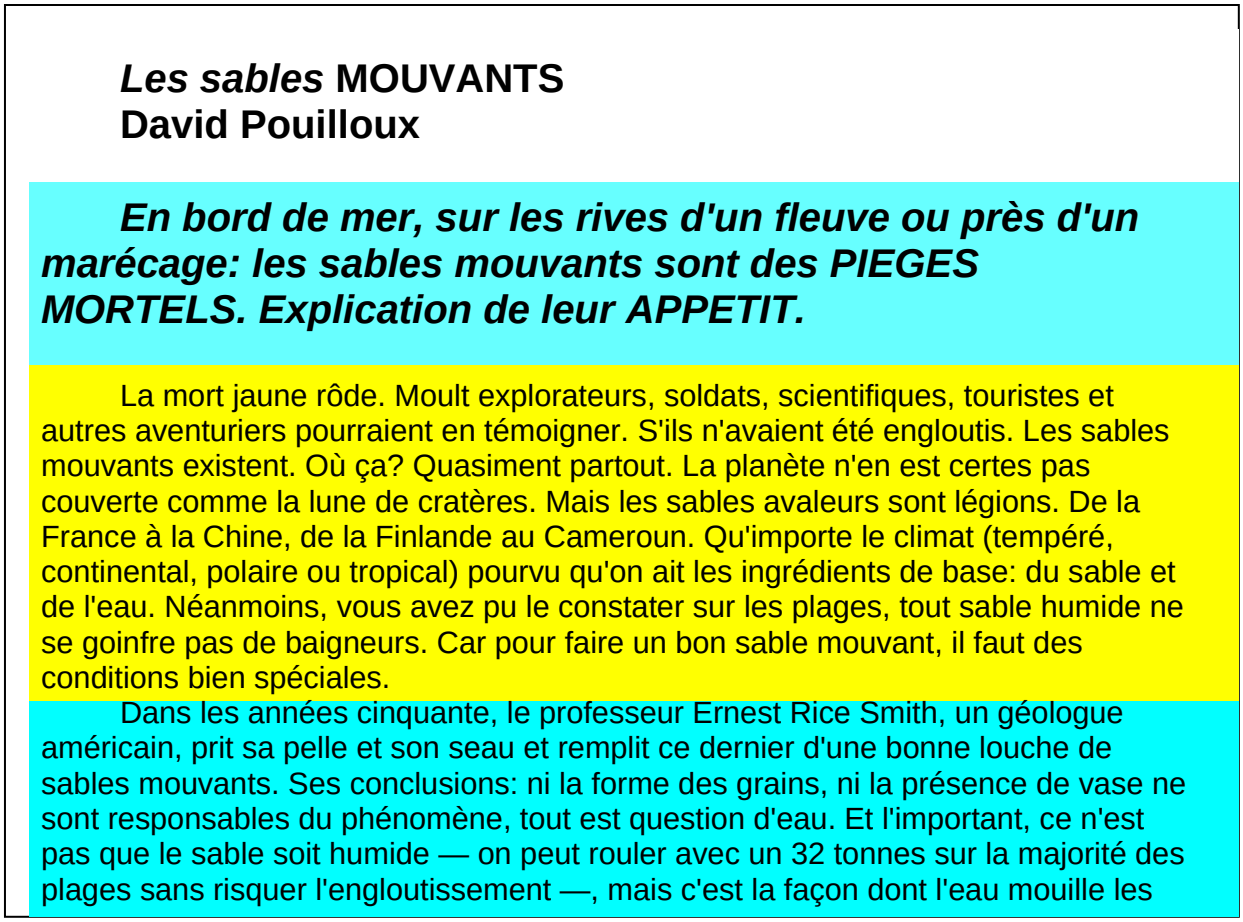
Les scripts et les commentaires dans le texte ne sont pas pris en compte. Nous les neutralisons à l'aide d'expressions régulières.

L'étude du code source HTML va nous être utile sur deux points :

- ✓ Connaître l'organisation du corpus va nous permettre dans un premier temps de connaître l'unité typographique de base du texte (par exemple le paragraphe). Cela va nous permettre de déceler l'explication de premier niveau si elle existe.
  On segmentera le texte à partir de l'unité typographique de base en sous unité typographique jusqu'à l'unité virgulot.

- ✓ Dans les corpus certains éléments sont différents des autres. Par exemple un groupe de phrases est quelque chose qui va souvent se répéter alors qu'un titre, un sous titre, une légende sont des éléments qui ont des caractères inattendus car ils peuvent être uniques ou peu fréquents dans le texte. On va donc typer chaque élément de notre corpus en lui affectant un caractère spécial ou non.

Exemple :



***Les sables* MOUVANTS**
David Pouilloux

En bord de mer, sur les rives d'un fleuve ou près d'un marécage: les sables mouvants sont des PIEGES MORTELS. Explication de leur APPETIT.

La mort jaune rôde. Moult explorateurs, soldats, scientifiques, touristes et autres aventuriers pourraient en témoigner. S'ils n'avaient été engloutis. Les sables mouvants existent. Où ça? Quasiment partout. La planète n'en est certes pas couverte comme la lune de cratères. Mais les sables avaleurs sont légions. De la France à la Chine, de la Finlande au Cameroun. Qu'importe le climat (tempéré, continental, polaire ou tropical) pourvu qu'on ait les ingrédients de base: du sable et de l'eau. Néanmoins, vous avez pu le constater sur les plages, tout sable humide ne se goinfre pas de baigneurs. Car pour faire un bon sable mouvant, il faut des conditions bien spéciales.

Dans les années cinquante, le professeur Ernest Rice Smith, un géologue américain, prit sa pelle et son seau et remplit ce dernier d'une bonne louche de sables mouvants. Ses conclusions: ni la forme des grains, ni la présence de vase ne sont responsables du phénomène, tout est question d'eau. Et l'important, ce n'est pas que le sable soit humide — on peut rouler avec un 32 tonnes sur la majorité des plages sans risquer l'engloutissement —, mais c'est la façon dont l'eau mouille les

Fig.2 Exemple de texte segmenté en parties : on remarque bien que le titre n'a pas la même mise en forme que le reste du texte.

Plusieurs solutions ont été envisagées pour réussir à découper le texte comme on le souhaite à partir du balisage HTML. Nous allons vous présenter ces différentes solutions et vous indiquer la solution retenue.

4.2 Première méthode : chercher pour chaque balise de début sa balise de fin.

Exemple : <P>texte 1<I> texte2</I></P>

Dans cette solution, qui se base sur la reconnaissance de la balise fermante, on aurait fait une segmentation incorrecte car en effet des balises ouvrantes ne sont pas forcément fermées ou bien ne sont pas fermées au bon endroit.

Et si l'on se retrouve dans le cas ci-dessous nous ferions une mauvaise fragmentation :

Exemple : <I><P>du texte en italique</I>du texte en style normal</P>

En effet nous n'aurions pas pris en compte le paragraphe mais une unité plus petite (le texte entre les balises <I></I>). Notre fragmentation n'aurait pas été correcte.
Le principe doit être de partir du « gros grain » vers un grain de plus en plus fin.

L'autre inconvénient est que le parseur peut prendre des trop gros morceaux de texte.

Exemple : <FONT COLOR='blue'><B>titre</B>
<P>texte</P><FONT COLOR='red'>texte en rouge</FONT>

Ici on aurait pris tout le texte et on ne se serait pas arrêté sur la détection du titre car la balise </FONT> a été omise après </B>.

La méthode a donc été abandonnée après différents tests sur des corpus. Elle est valable sur du HTML bien formé.

4.3 Deuxième méthode : travail sur toutes les balises

Celle-ci se base aussi sur la recherche des balises ouvrantes et fermantes mais on explore directement toutes les balises à l'intérieur de notre espace de recherche.

Exemple : <P><B>texte gras</B>texte normal<I>texte en italique</I></P>

Dès que l'on rencontre une balise ouvrante on cherche à voir si l'on peut en trouver une autre ouvrante pour aller chercher la mise en forme matérielle incluse.

Cette méthode est trop stricte par rapport au balisage mal formé et a été abandonnée rapidement. En effet comme dans la méthode précédente si on ne trouve pas la balise fermante le parseur « se perd ».

4.4 Troisième méthode : recherche des balises fermantes collées à une ouvrante

Cette méthode s'occupe des balises fermantes.

En effet nous nous arrêtons lorsque nous trouvons des balises fermantes suivies de balises ouvrantes mais nous ne regardons pas le type de balise dans un premier temps (P, TD, I… peu importe).

De ce fait une balise ouvrante peut être oubliée d'être fermée.

Exemple : <P>texte 1<I>texte 2 <B>texte3</I></B><P>suite…

En fait cet algorithme permet de « manger » le maximum de balises fermantes et avec ce procédé on va toujours repartir avec une balise ouvrante en début du reste du texte à traiter.

Avec cet algorithme j'ai pu détecter les parties du texte et donc l'unité typographique de base.

L'expression régulière utilisée est la suivante :

```
$expReg="#(<[A-Z|a-z])(.*?)(>([\[|\w|[:blank:]](.*?)?(</(.*?))|(</(.*?))))(?:<[A-Z|a-z|[:blank:]|.])#";
```

L'algorithme, une fois lancé sur un texte, va stocker les différentes parties trouvées dans une table de notre base de données. C'est à partir de cette table que nous allons travailler pour la suite des étapes.

4.5 Un algorithme retenu tolérant mais qui ne résout pas tout

En effet cet algorithme retenu nous a paru être le plus efficace après l'avoir tester sur de nombreux textes.

Cependant il ne fonctionne pas correctement avec les documents générés par certaines versions de Microsoft Word qui génère beaucoup de « namespaces » qui rendent le parsing difficile.

Certains de ces textes ne s'affichent pas dans les navigateurs.

Dans ce cas, on peut faire un prétraitement sur le code source HTML avec des utilitaires comme TIDY, qui vont permettre de nettoyer le code. Le résultat n'est malheureusement pas celui attendu et on se retrouve avec un corpus qui n'est pas vraiment exploitable pour notre algorithme.

Cette dernière méthode ne permet pas de travailler avec des textes qui ne comportent pas de balises fermantes. En effet, j'ai certains textes qui ne possèdent aucune balise fermante.

TIDY peut y remédier, et le texte sera pris par notre parseur.

Le travail fait jusqu'ici est uniquement un découpage du texte à « gros grain ».

Une fois cette étape effectuée, il a fallu caractériser les parties pour savoir si elles avaient un caractère spécial ou non par rapport au reste du texte.

5 Catégorisation des parties

Ce travail nous permet de voir si une partie de texte à un caractère inattendu par rapport au reste du texte. En règle général, le titre, le pied de page ont des mises en forme différentes que le reste du texte.

Cette classification va se faire par rapport à trois critères :

- ✓ L'unique, le multiple
- ✓ La position des marques entre elles
- ✓ La longueur

5.1 L'unique et le multiple

Une marque qui se trouve une seule fois dans le texte possède un caractère inattendu.

En effet, ici on se base sur la différence entre l'unique et le multiple.

C'est une mise en forme matérielle que l'on retrouvera une seule fois dans le texte. La partie titre d'un document est, en général, unique au niveau mis en forme matérielle dans un texte.

Exemple :

1999-04-27:
FRANCE: ECONOMIE - DANS LA FAMILLE AIRBUS, L'A318 EST NÉ. L'AVION DE CENT PLACES VOLERA EN 2002. ***(LIBRTN)***

Un Airbus tout petit viendra agrandir la famille des avions européens. Le consortium Airbus Industrie a annoncé officiellement hier le lancement de son plus petit avion, l'A318

...

Fig.3 La mise en forme du titre est bien unique et inattendue par rapport au reste du texte

Ici on remarque bien que le titre est différent du reste du texte au niveau de sa mise en forma matérielle.

5.2 Critères sur la position des marques

Le critère de position des marques est très important et primordial pour le typage de nos parties.

En effet, si l'on rencontre deux fois une même mise en forme matérielle pour deux parties de texte différent, cela ne signifie pas que ces passages de texte sont « ordinaires ». On ne peut pas se baser sur le seul critère de l'unique et du multiple.

Par exemple, on peut retrouver des informations en sous-titre ou en chapeau qui ont même mise en forme matérielle qu'une partie du pied de page du texte (après le corps même du texte, comportant par exemple l'auteur, les références…). Ces deux parties ont bien des

caractères inattendus par rapport au reste du texte car tous deux marquent une partie bien spéciale du texte.

Donc on peut dire que deux mises en forme matérielles identiques mais éloignées les unes des autres sont des marques spéciales. Le critère d'éloignement se fait par rapport à la disposition générale de chaque partie entre elles (en moyenne une partie par rapport à une autres est espacé par une longueur x).

5.3 Le critère de longueur

Une partie de texte courte et isolée peut être une marque spéciale.

Là aussi le critère de longueur fait référence au reste du texte.

Si on se situe au niveau de la phrase notre référence pour la longueur sera évidemment la longueur moyenne d'une phrase.

Exemple :

Engendrées par le vent, les vagues qui se développent dans la zone où le vent souffle et qui peuvent se propager hors de cette zone, sont appelées ondes de gravité. Elles possèdent des caractéristiques spatiales et temporelles bien précises. Les paramètres d'une vague de forme sinusoïdale, forme épurée et théorique, sont la hauteur entre la crête et le creux, et la longueur d'onde, à savoir la distance qui sépare deux crêtes successives. La grandeur caractéristique du temps est la période : placez-vous en un point fixe par rapport au fond, vous voyez défiler chaque vague. Le temps qui sépare le passage de deux crêtes successives est la période de la vague.

Brisant de récif, à Hawaï.

…

Fig.4 Un texte court peut être une maque spéciale

On remarque bien dans ce texte, qu'en plus de sa mise en forme matérielle en gras, on a un texte court qui est « spécial » par rapport au reste du texte.

Avec tous ces critères nous avons pu catégoriser les parties de plus haut niveau (« gros grains ») que nous avions trouvé.

Ces techniques vont nous resservir aussi dans la suite pour détecter de la même manière les parties de texte inattendues quand on va descendre d'un pas dans notre hiérarchie d'unité typographique (…, phrases, virgulots,…).

Maintenant nous allons pouvoir détecter le corps du texte à partir du travail que l'on vient d'effectuer.

6 Détection du corps de texte

Le corps de texte est très important car c'est dans cet espace que va se situer une grande partie de notre recherche.

C'est dans cette partie que l'on détectera des indices pouvant nous révéler que le texte a une valeur explicative ou pas (en premier lieu on regarde le titre, nous y reviendront plus bas).

Cette étape va nous permettre d'extraire le corps du document mais aussi ce qu'il y a avant c'est-à-dire le titre.

On a vu, plus haut dans le rapport, que le corps du document était formé globalement de sections ordinaires (on regarde ici à « gros grains ») et que le titre était une marque spéciale de par sa mise en forme matérielle, qui différait du reste du texte.

On peut donc dire qu'un texte est un document composite formé d'une partie spéciale qui va être le titre, sous-titre, chapeau…, d'une partie corps de texte de forme « ordinaire » puis d'une partie pied de page (si elle existe) qui sera aussi de forme spéciale.

Exemple :

Le point chaud de l'Afar sous surveillance

Près de 90% des volcans naissent en bordure des plaques tectoniques, au niveau des dorsales et des plaques de subduction. Mais il existe un deuxième type de volcanisme, beaucoup moins répandu, dont l'origine ne semble pas être liée aux mouvements tectoniques : le volcanisme de point chaud. " *Certains volcans apparaissent au milieu des plaques lithosphériques et résultent de la remontée rapide de matière chaude provenant des profondeurs du manteau,* explique Jean-Paul Montagner, directeur du Département de sismologie de l'Institut de physique du globe de Paris (IPGP).

…

…

Fin 2001, nous devrions être en mesure de fournir une image détaillée du sous-sol de la corne africaine."

Jacques Gozzo

Contact: **Jean-Paul Montagner**.

Département de sismologie IPGP, UMR 7580, Paris.

Fig.5 Les différentes parties d'un texte

Quand nous allons passer à l'échelle du corps du document nous allons aussi retrouver un document composite car on travaillera avec un grain plus fin.

Cette diversité dans les formes matérielles dans le corps du document nous permettra de détecter des marqueurs pour des explications ou bien renforcer ce que les indices typographiques nous ont décelé.

Maintenant que nous avons détecté les différentes parties de notre texte nous allons pouvoir découper le texte avec une unité typographique plus petite.

7 Segmentation du corps de texte

La segmentation du corps de texte a été faite avant de travailler sur la recherche d'explications en elle même afin de ne pas ralentir le traitement de cette recherche.

La segmentation consiste à découper notre texte en unité typographique plus petite que l'unité typographique de base.

Nous arrêterons notre segmentation au niveau du virgulot.

7.1 Stockage de l'information dans une base de données

Pour chaque segmentation de niveau n-1, n-2… (n étant le niveau de l'unité typographique de base) je vais créer une table dans ma base de données.

Dans ces tables seront stockés des informations concernant le contenu du texte, ainsi qu'une référence sur le père du segment découpé.

Par exemple, dans le cas du paragraphe comme unité de base du texte, on sait qu'une phrase appartient à un paragraphe et qu'un virgulot appartient à une phrase.

On stocke des informations sur la position des segments (par rapport à son père).

On pourra donc savoir que la phrase 411 sera la 4ème phrase du 3ème paragraphe par exemple.

On a donc pour chaque segment soit sa position relative (au segment père) soit sa position absolue.

On stocke aussi des informations sur la longueur du segment extrait.

La longueur permettra de travailler sur l'offset, si on le souhaite, ou encore caractériser les segments trouvés avec les critères vus dans la partie 4 (unique / multiple / position / longueur).

Niveau_n
Id
Pere
Pos
Longeur
Texte

Après toutes ces étapes nous avons à notre disposition :

- ✓ Le texte d'entrée
- ✓ Ses différentes parties (titre / corps / pied de page)
- ✓ L'unité typographique de base
- ✓ n tables contenant le texte segmenté sur n niveaux.

Nous allons maintenant travailler maintenant sur la détection d'explications dans le texte.

8 Le couple « Posé » / explication

8.1 Présentation

Une explication est toujours introduite par une partie qui va poser le problème, qui va amener la question à laquelle on attend une réponse. Nous nommons donc cette partie le « posé ».

Le « posé » est unique alors que l'explication peut être multiple.

En effet, on pose le problème puis on peut y répondre par un développement qui va être divisé en parties et chaque partie sera une explication pour le « posé ».
On doit donc chercher les formes factorisées qui vont avoir une valeur explicative pour le problème posé.

8.2 Détection

8.2.1 Le posé

Le posé a une valeur inattendue car il est unique. Cette forme inattendue peut être détectée soit par la mise en forme matérielle du texte, soit par la recherche d'indices syntaxiques spécifiques.

Nous avions travaillé jusqu'à présent sur la mise en forme matérielle du texte. On va pouvoir étudier dans les parties suivantes le travail sur les indices morphologiques.

Un posé peut être introduit par une question ou une négation.
En effet à la suite d'une négation on s'attend à avoir une justification de cette négation.

Exemple : Le gouvernement n'a pas su réagir face ….

On s'attend à trouver une argumentation expliquant pourquoi le gouvernement n'a pas su réagir face a tel ou tel problème.

La fin du posé va être marqué par le début de l'explication.

8.2.2 L'explication

Elle aussi, elle est marquée par des marques spécifiques.
On va la repérer en mettant en relation un couple de marques qui vont nous permettre de délimiter cette explication.
Une explication suit généralement un « posé ».

9 Structure d'un texte explicatif

Un texte explicatif, vu à gros grain, peut avoir deux formes au niveau de sa structure :

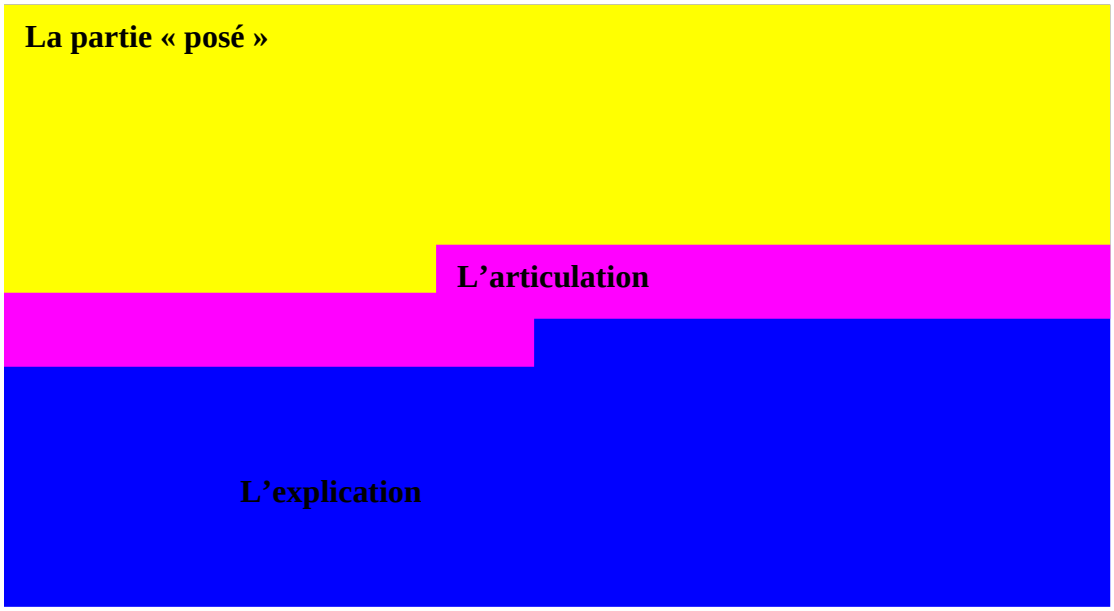
- ✓ Soit une structure binaire où l'on retrouve un « posé », une articulation « centrale » et une explication.
- ✓ Soit une structure d'enchâssement

9.1 La structure binaire

La structure binaire possède une articulation centrale qui fait la liaison entre le « posé » et l'explication.

A l'intérieur de la section contenant l'explication, on peut aussi retrouver une structure d'enchâssement.

Schématisation :



9.2 La structure d’enchâssement

La structure d’enchâssement a une forme différente car on a un « posé » puis une explication pour ce « posé ». A l’intérieur de cette explication on peut retrouver des sous explications qui expliquent le « posé » (c’est la factorisation qui a été abordée plus haut).

Schématisation :

Pour détecter la structure d’un texte nous verrons un peu plus loin qu’un travail sur le titre peut nous apporter des éléments utiles.

Cette recherche d’information, dans le titre comme dans le corps du texte, en rapport avec notre recherche de couple « posé / explication », va se faire, comme nous l’avons dit, par la recherche d’indices spécifiques.

Nous allons voir comment nous détectons ces indices et comment nous les exploitons pour trouver nos passages de texte explicatifs.

10 Les marqueurs

10.1 Présentation

Nous utilisons un ensemble de marqueurs pour détecter des débuts ou des fins d'explications ou de posés.

Chaque marque va avoir un type et une portée.

Extrait du tableau des marques :

Id	Description	Expression régulière	type
16	Voilà comment	(Voilà\scomment)	phrase explicative
17	Qui ... ?	(Qui(.*?)?(\?))	phrase interrogative
18	n' ... pas	\sn\'(.*?)?pas\s	proposition négative
19	Comment... ?	(Comment(.*?)?(\?))	phrase interrogative
22	Quel...?	(Quel(.*?)?(\?))	phrase interrogative
21	Pourquoi...	(Pourquoi(.*?)?(\?))	phrase interrogative

Certaines marques ne vont être visible que dans une certaine unité typographique.

Exemple : *« Qui n'a pas été fasciné par le spectacle des vagues, se faisant et se défaisant, répétitif, parfois violent et chaotique ? »*

Si on cherche les interrogations, on ne pourra pas les détecter intégralement au niveau du virgulot.

Si on se situe au niveau du virgulot on ne pourra pas détecter des questions de la forme « Qui … ? ». Le mieux que l'on puisse faire ici, au niveau virgulot c'est de détecter le « ? » mais cela aura déjà été fait lors du travail sur l'unité phrase.

Les marqueurs sont détectés à l'aide d'expressions régulières.

A chaque fois que nous détectons un marqueur nous insérons une marque dans notre texte.

Cette marque qui est apposée possédera les propriétés suivantes :

- ✓ Son type : le type de marque : interrogative, explicative…
- ✓ Sa portée : l'unité typographique sur laquelle on travaille.
- ✓ Son numéro : son numéro dans l'unité typographique où l'on se trouve (est-ce la première marque ? la deuxième ?)
- ✓ Sa référence : identifiant du segment sur lequel elle a été trouvée.

Nous avons donc un texte qui est marqué par des balises, qui font références à des indices dans une unité précise.

Nous allons maintenant voir que c'est la mise en relation de ces marques qui vont permettre de détecter nos couples « posé / explication ».

10.2 Leurs relations

10.2.1 Notion d' « appel / echo »

Cette notion permet de lier deux éléments entre eux et de donner un sens à cette liaison.
On pourra considérer un élément de la liaison comme « appel », noté A et l'autre « echo » noté E.
A et E vont délimiter un intervalle dans le texte et cet intervalle pourra servir pour détecter une explication par exemple. A et E sont des formes syntaxiques dans le texte.

Exemple de liaisons « appel-echo »:

Appel, A	Echo, E
phrase définitoire	phrase conclusive
phrase interrogative	phrase explicative
phrase négative	phrase conclusive
proposition négative	phrase conclusive
proposition définitoire	proposition conclusive

En plus de la relation qu'il peut exister entre deux marques, il va falloir étudier la position de ces marques les unes par rapport aux autres.
C'est ce qui va nous permettre de savoir si un segment délimité par deux marques est un « posé » ou une explication.
Par exemple nous savons qu'un « posé » donne généralement suite à une explication.
Donc pour une unité typographique « m », si l'on trouve un premier couple de marqueurs, il s'agira du « posé » pour ce niveau.

Nous venons de présenter les techniques que nous employons pour détecter nos explications, nous allons maintenant voir comment nous mettons en œuvre ces techniques pour en sortir un texte analysé.

Rappel :

- ✓ Nous possédons en texte segmenté de l'unité typographique de base jusqu'au virgulot.
- ✓ Nous avons détecté trois parties : zone de titre, corps de texte, zone de pied de page.
- ✓ Un texte peut avoir une structure binaire ou d'enchâssement.
- ✓ Nous avons défini des marques à détecter pour chaque unité typographique et nous les avons repéré dans le texte.

- ✓ Nous travaillons sur la relation entre des marques pour définir nos couples « posé / explication ».
- ✓ La position des marques est importante.
- ✓ Il faut tout de même chercher si le texte a une valeur explicative avant sur les premiers niveaux avant de descendre dans des unités typographiques plus petites.

11 Travail sur le titre

Le titre va nous permettre de nous éclairer sur trois points :

- ✓ La valeur explicative du texte.
- ✓ La structure du texte.
- ✓ L'espace de recherche pour le « posé » de premier niveau dans le corps de texte.

11.1 Une aide pour la valeur explicative du texte

En effet le titre va nous permettre de donner des indications sur la valeur explicative d'un texte ou tout du moins de savoir si le corps du texte a valeur d'explication pour le titre.

Nous regardons donc la partie de titre du document (niveau 1 pour l'unité typographique). La partie de titre doit contenir au moins deux éléments pour valider ces règles.

- ✓ Si la partie du titre possède un caractère interrogatif ou négatif et que le dernier segment du corps est résultatif alors le corps est une explication pour la partie titre.

- ✓ Si le titre est neutre et que le dernier segment est une subordonnée résultative alors le corps est une explication pour la partie titre.

On peut donc détecter, pour certains cas, pour le premier niveau de recherche, la valeur explicative du corps de texte pour le titre.
On pourra donc rechercher des explications à l'intérieur de cette partie détectée (partie délimitée par le couple A-E).

11.2 Le titre, un indice pour la structure du document

En utilisant les mêmes règles que pour la détection de la valeur explicative du corps de texte on pourra détecter, dans certains cas, la structure du texte :

- ✓ Si la partie du titre possède un caractère interrogatif ou négatif et que le dernier segment du corps est résultatif alors la structure du texte est une structure d'enchâssement.

- ✓ Si le titre est neutre et que le dernier segment est une subordonnée résultative alors la structure du texte est une structure binaire articulée.

11.3 Le titre nous aide à délimiter le posé de premier niveau dans le corps de texte

Dans le cas où la section de titre est composée de plusieurs partie niveaux 1, alors cela nous donne une indication sur l’espace de recherche pour le « posé » de premier niveau du corps de texte.

Si la partie titre comporte un titre, un sous titre et un chapeau alors l’espace de recherche pour le « posé » de niveau 1 se fera sur les trois premières parties du corps de texte (en rapport avec l’unité typographique de base).

Exemple :

Dans le texte ci-dessous en rouge la partie de titre et en jaune l’espace de recherche pour le premier niveau d’explication dans le corps de texte. En Mauve le « posé », qui est bien inclus dans notre espace de recherche.

La mécanique des vagues

Des lames tubulaires d'Hawaï à la puissante barre de Grand Bassam en Côte d'Ivoire, les vagues semblent échapper à toute description figée. Elles obéissent pourtant à un cycle de vie identifiable.

Par Loys Schmied

Qui n'a pas été fasciné par le spectacle des vagues, se faisant et se défaisant, répétitif, parfois violent et chaotique ? L'image symbolique des vagues est la passivité, passivité de la houle qui festonne le fluide sans un souffle de vent, passivité dans la violence, la puissance, l'inertie qui se communique à la masse fluide. Une vague est une onde qui se propage à la surface de l'eau. Comme toutes les ondes, les vagues transportent de l'énergie, mais ne transportent pas de matière. La masse d'eau, mise en mouvement au passage d'une vague près du'rivage reste près du rivage. De même, la masse d'eau, mise en mouvement au large par cette même onde, reste au large. L'image du drap que l'on secoue de haut en bas en est une bonne illustration : si vous secouez fortement un drap à une extrémité, vous engendrez des ondulations qui se propagent à l'autre extrémité. Vous pouvez suivre la déformation du tissu qui se propage d'un bout l'autre du drap, mais la matière, le coton du drap, n'est pas deplace.

Mais lorsqu'une vague déferle, n'y-a-t-il pas transport d'eau ? Certes, mais il s'agit de déplacements locaux dont les dimensions sont sans commune mesure avec la distance de propagation d'une onde, qui peut atteindre plusieurs centaines de kilomètres.

Engendrées par le vent, les vagues qui se développent dans la zone où le vent souffle et qui peuvent se propager hors de cette zone, sont appelées ondes de gravité. Elles possèdent des caractéristiques spatiales et temporelles bien précises. Les paramètres d'une vague de forme sinusoïdale, forme épurée et théorique, sont la hauteur entre la crête et le creux, et la longueur d'onde, à savoir la distance qui sépare deux crêtes successives. La grandeur caractéristique du temps est la période : placez-vous en un point fixe par rapport au fond, vous voyez défiler chaque vague. Le temps qui sépare le passage de deux crêtes successives est la période de la vague.

Brisant de récif, à Hawaï.

Sur la côte nord de Hawaï, du fait de l'absence de plateau continental, les lames heurtent les récifs de l'île avec la puissance quasi intacte du plein océan.

Anatomie d'une vague.

Une lame est caractérisée par sa longueur d'onde, la distance horizontale séparant deux crêtes ou deux creux successifs, sa hauteur, la distance verticale entre le sommer de la crête et la base du creux, et sa période, le temps mis par une crête pour parcourir une longueur d'onde.
…
…

Nous possédons à présent tous les éléments pour construire notre algorithme de détection. Nous allons vous le présenter dans la partie qui suit.

12 Algorithme de reconnaissance et mis en relation

Nous travaillons d'abord au premier niveau de l'article en examinant le titre et le corps du texte. Je vais travailler sur les ponctuations et les mots spéciaux.

1) Je regarde si le titre est marqué.

2) Je regarde si le corps est marqué.

a) Si le corps et le titre ne sont pas marqués alors je dis que le texte n'est pas explicatif

b) Si a) n'est pas vérifié

i) Si le titre est marqué et si le segment de fin du corps de texte est marqué alors je colorie tout le corps de texte. Cela signifie que le corps de texte est une explication pour le titre. Le titre sera donc le posé de premier niveau.
Je passe à l'étape 3.

ii) Si le titre est neutre et que le corps est marqué alors je cherche dans le corps de texte un couple « posé / explication » pour l'unité typographique de recherche en cours, je passe donc à l'étape 3.

3) Détection d'un couple « posé / explication »

Je rentre donc dans le corps de texte, avec l'unité typographique de niveau n.
Je regarde les parties de ce niveau qui sont marquées et je les mets en relation en regardant les relations que j'ai dans la table « relations ».
En fonction de ces relations je détecte donc les couples « posé / explication ».
Ensuite je repasse a cette étape avec une unité plus petite et ce jusqu'au virgulot pour détecter les explications plus locales.

13 Etude de textes en anglais

Les textes en anglais sont des textes de langue latine. Le travail sur la détection des parties, sur la segmentation reste inchangé.

L'algorithme de détection des couples « posé / explication » lui aussi est le même, seuls les marques de détection vont changer.

Les marques spécifiques pour l'anglais ne sont pas actuellement exhaustives dans la table des marques.

14 Problèmes rencontrés

Le balisage est évidemment le problème majeur quand l'on travaille sur un document HTML. Nous avons vu que notre méthode pour le travail sur le code HTML n'accepte pas tous les textes en entrée.

La détection du corps de texte n'est pas facile et notamment concernant la détection du pied de page. En effet en se basant sur les critères d'unique, multiple, longueur et mis en forme matérielle pour le typage des parties nous pouvons faire des erreurs.

Il faut aussi détecter des caractères spéciaux tel que le « © » qui se retrouvent souvent dans le pied de page d'un article.

15 Conclusion

Ce travail m'a permis de voir, une fois de plus, que le travail d'analyse est une phase très importante lors de la réalisation d'un travail.

Je me suis aperçu que le travail des linguistes est un travail très complexe car une multitude d'éléments peut interagir et chaque interaction peut donner lieu à un nouveau critère.

Mais il faut réussir à déterminer des méthodes généralistes pour un ensemble de critères et non faire une règle pour un critère.

En effet, si on se base sur tous les critères, nous aurions des algorithmes lourds à manipuler et surtout fragiles car très spécifiques.

Ce travail réalisé utilise, je pense, quelques « briques » de la linguistique et peut être affiné en faisant une étude plus approfondie sur le corpus, ce qui améliorera notre algorithme de mis en relation des marques.