TAL : Annotation sémantique et ressources linguistiques

Cours 11

Master LFA, 2011/2012

Annotation sémantique. Applications

Annotation textuelle

- Dans *Le Robert* :
  - ***annotation*** : note critique ou explicative qui accompagne un texte
  - ***annoter*** : accompagner un texte de notes critiques ; mettre sur un livre des notes personnelles

- En informatique :
  - une ***annotation*** est un commentaire, une note, une explication ou tout autre remarque externe qui peut être attachée à un document ou à une partie de celui-ci.
  - l'annotation textuelle consiste à ***enrichir un texte*** avec des informations, rattachées aux parties du texte.

Annotation (en informatique)

- Le mot ***annotation*** signifie à la fois :
  - le processus dans lequel le texte est enrichi avec des informations supplémentaires.
  - les informations qui sont rajoutées au texte.

- Pendant l'annotation certains éléments textuels (mots, expressions, phrases, …) sont étiquetés par un ensemble de ***catégories d'annotation,*** suite à une analyse des propriétés de l'élément annoté selon une méthode donnée.
  - L'ensemble ***d'étiquettes*** ou de ***catégories d'annotation*** doit être défini préalablement.
  - Les annotations expriment certaines propriétés des éléments textuels, par ex. des catégories grammaticales, le contenu sémantique, etc. suivant un modèle.

Pourquoi annoter ?

- L'annotation a pout but *d'expliciter (ou de « traduire ») certaines propriétés des éléments textuels* (notamment le sens) qui sont normalement inaccessibles pour la machine.
- L'annotation ajoute une information au texte qui est *lisible à la fois par l'être humain et par la machine.*
- Elle permet donc à la machine *d'accéder au sens par le biais des annotations* (les étiquettes attribués aux éléments textuels) qui reflètent, si elles sont correctes, une partie du sens du texte.

Exemple 1 : analyse morpho-syntaxique

- Voir TreeTagger.
  - Les éléments annotés sont les mots.
  - Les catégories d'annotation sont des catégories morpho-syntaxiques (verbe, nom, adjectif, …) qui sont issues de l'analyse morpho-syntaxique en linguistique.

Exemple 2 : annotation automatique

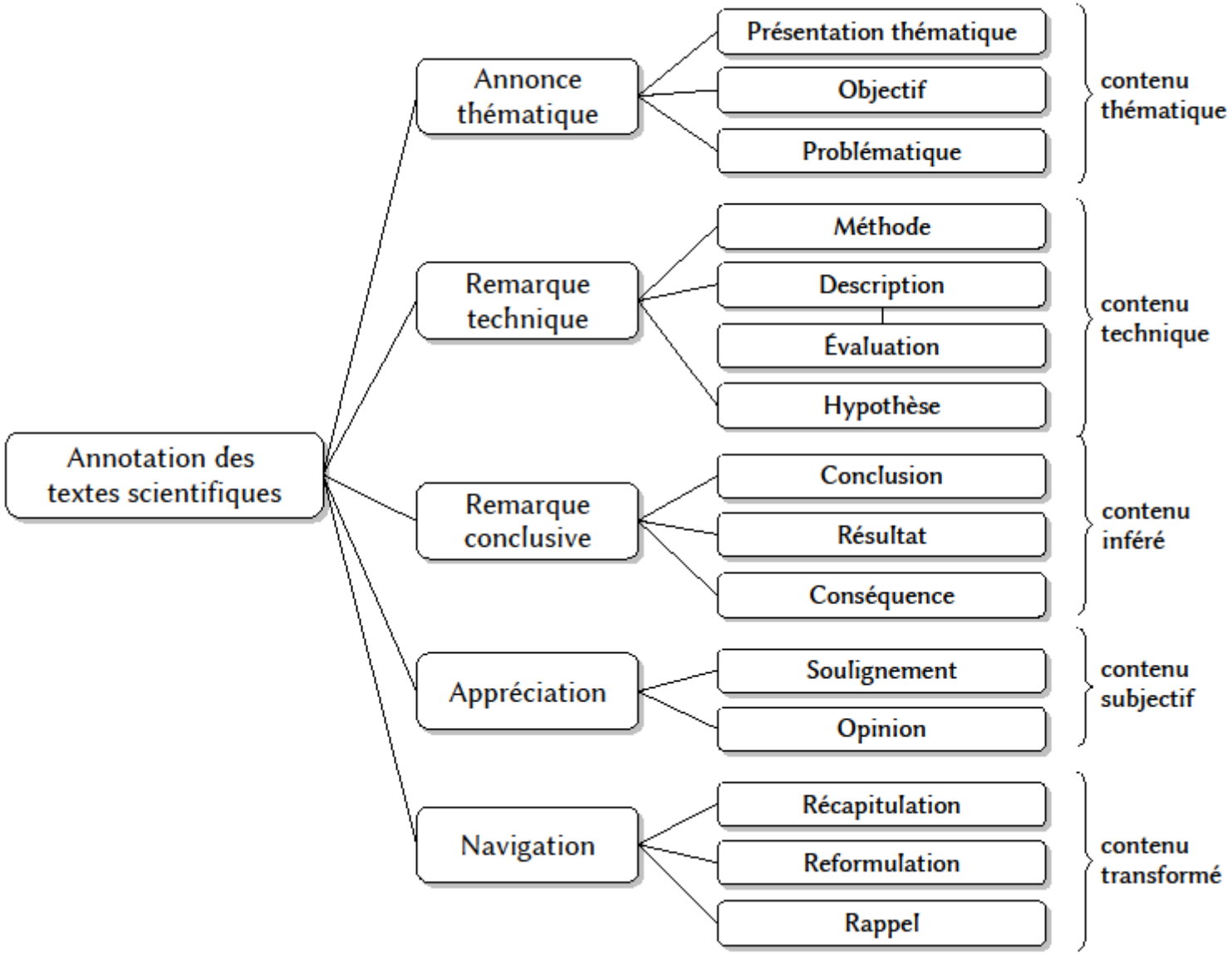
La recherche en ce domaine et, plus précisément, dans celui se consacrant à la communication médiée par les systèmes informatiques (Computer Mediated Communication, CMC), a particulièrement étudié la conversation textuelle par clavardage. (problématique) Ces travaux permettent, avec le recul du temps, de se faire une idée précise des possibilités offertes par ce moyen de communication directe et synchrone [Calico05] . (résultat) Mais, même si un faisceau convergeant de résultats montre l'impact que peuvent avoir certaines formations qui en font largement usage sur les compétences orales, ce n'est que depuis peu que des environnements technologiques offrent la possibilité de mettre les apprenants en situation simultanée de compréhension et production orales dans de vrais dialogues, ou plutôt polylogues. Les premières recherches rassemblant audio et clavardage mettent ces deux modalités de communication en concurrence [Sykes05] ou se limitent à des échanges individuels apprenant-enseignant [Blake05]. (résultat) Nous nous intéressons ici à l'usage combiné de ces modalités dans des échanges de groupe et, plus encore, à l'étude de leur association à d'autres modes et modalités dans des espaces en ligne où se construisent des discours d'un genre nouveau. (objectif) La compréhension de l'agencement de cette multimodalité est sans doute la clef pour le passage à un enseignement complet (dans les quatre compétences) des langues en ligne.

Annotation par A. Blais, 2008

Annotation automatique

- Dans l'exemple précédent :
  - les éléments annotés sont les phrases
  - les catégories d'annotation sont un certain nombre de catégories discursives (*problématique, résultat, objectif, …)* qui s'inscrivent dans une ***ontologie*** (*voir diapos suivantes)*
  - Il n'existe pas de théorie linguistique (assez stable et accepté par la communauté) qui permet de catégoriser les phrases d'un texte argumentatif.
  - L'ontologie ici a été créée spécifiquement pour répondre à un besoin précis.

Exemple : ontologie linguistique des catégories discursives dans des textes scientifiques *(A. Blais, 2008)*



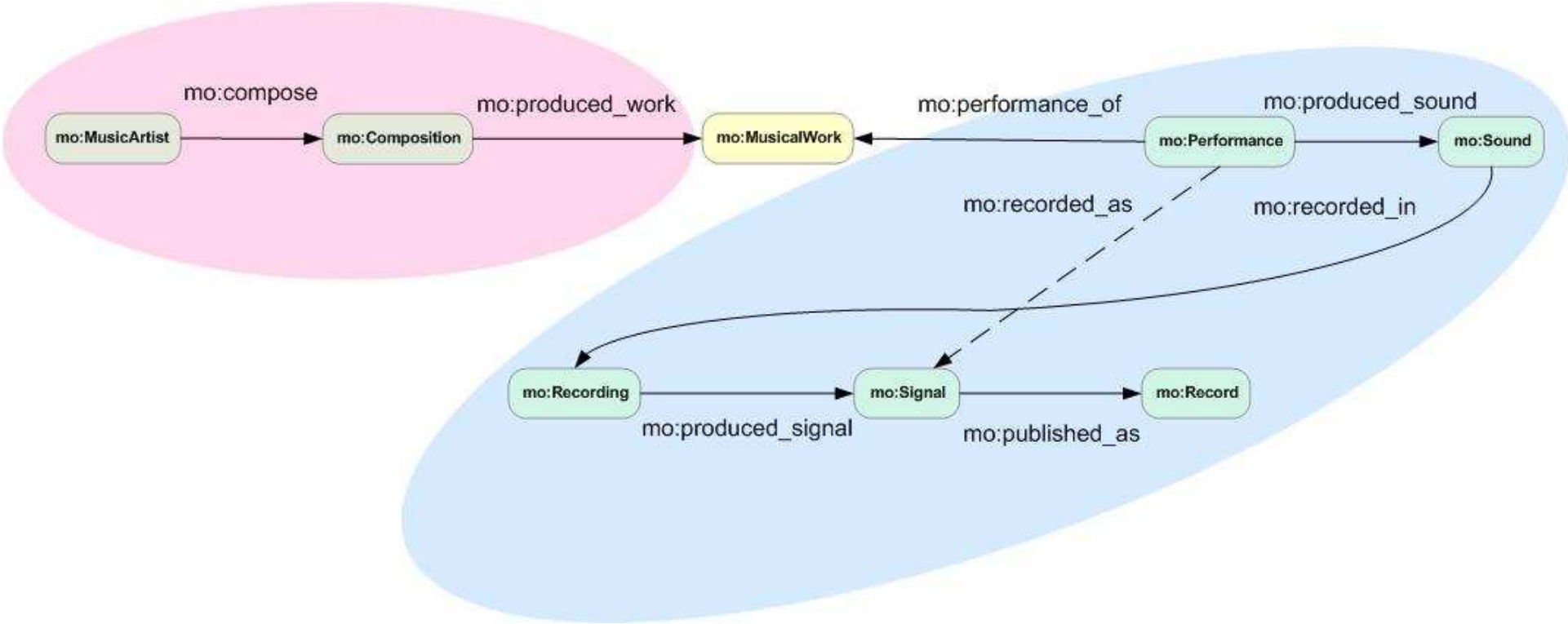
Ontologie

- En philosophie :
  - ***L'ontologie*** est l'étude de l'être en tant qu'être, c'est-à-dire l'étude des propriétés générales de ce qui existe.

- En informatique :
  - ***Une ontologie*** est un ensemble structuré de concepts représentant le sens d'un champ d'informations. Les concepts sont organisés dans un graphe, où les relations entre eux peuvent être de différents types.
  - ***Ontologie de domaine*** : permet de modéliser un ensemble de connaissances dans un domaine donné.

Ontologie : exemple

- Exemple d'une ontologie avec différents types de relations entre les concepts.



Système d'annotation automatique

- Conditions nécessaires pour un système d'annotation :
  - déterminer quels sont les *éléments textuels* à annoter.
  - pouvoir identifier ces éléments *automatiquement* dans le texte !
  - déterminer l'ensemble des catégories d'annotation et les éventuelles relations entre elles, c'est-à-dire construire *l'ontologie*.
  - chaque catégorie d'annotation doit être l'objet d'une *définition formelle* : on doit être capable de dire pour n'importe quel élément s'il appartient à la catégorie ou non !
  - avoir un algorithme qui serait capable d'attribuer les catégories aux éléments textuels.

Annotation automatique

Segmentation

- identifier les sections, paragraphes, phrases, mots, … dans le texte : donc identifier les éléments qui seront éventuellement annotés

Annotation

- **Approche statistique :** calcul des fréquences, n-grammes, …
- ET / OU
- **Approche par ressources linguistiques :** examiner chaque élément dans son contexte, et vérifier des conditions

Ressources linguistiques

Evaluation

- Calcul des mesures rappel / précision

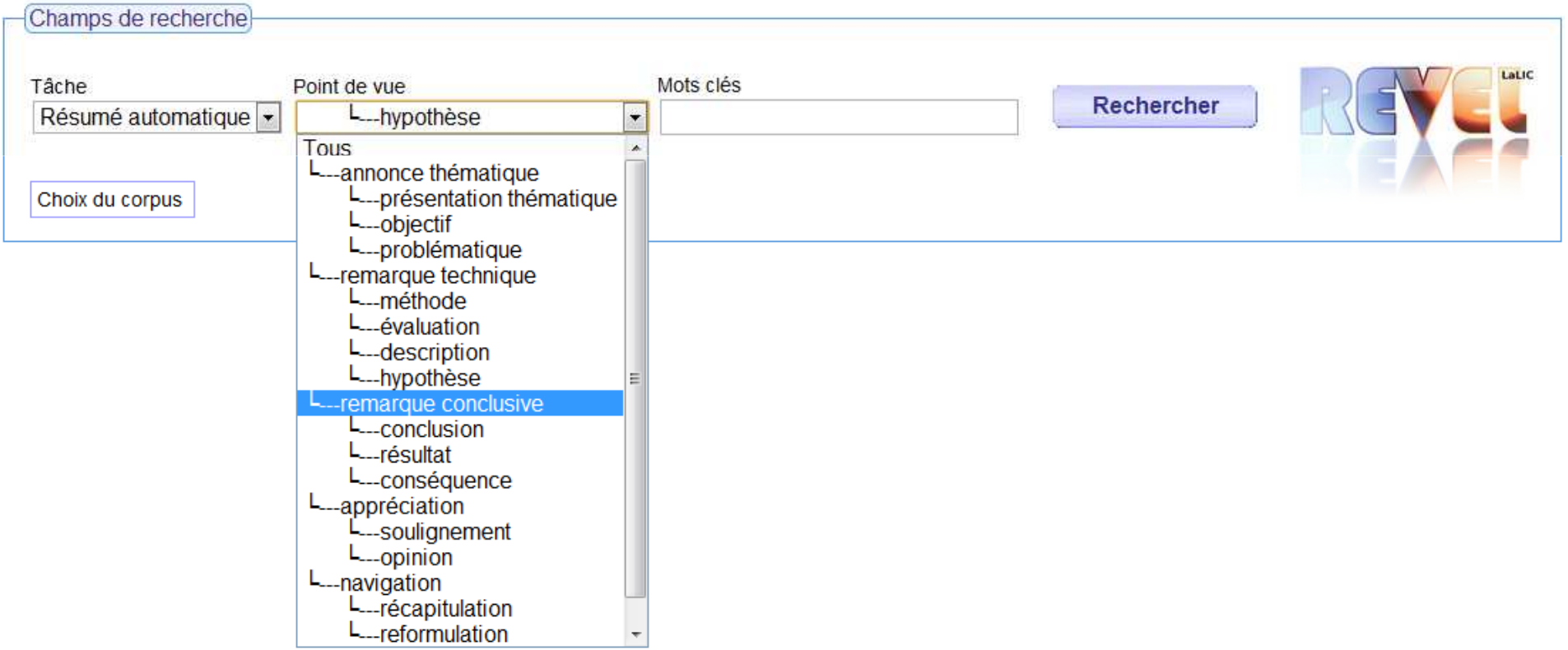
Applications de l'annotation sémantique

- Recherche d'informations
- Résumé automatique
- Extraction automatique des hypothèses, définitions, relations de cause à effet, …
- Fouille de textes juridiques
- Analyse des opinions (sur les forums, réseaux sociaux, …)
- Analyse des relations entre personnes (dans des publications scientifiques, dans des articles de presse)
- …

- L'annotation textuelle est une des étapes de presque toutes les applications en TAL.

Exemple : recherche d'information sémantique

- La recherche s'effectue à la fois sur des termes, mais aussi sur des « points de vue de fouille » ou des catégories discursives qui ont été annotées dans les textes.



Exemple : résultats

- Recherche des définitions, mot clé « *beauté* ».

Résultats

1. corpus.fr.utf8.t4t3t0.EcoleDoctorale3.these.Vigneron.txt.xml (Projet-theses, bibliosemantiqueETdefinition)
En tant qu'idée locale, la beauté est une princesse despotique, et sujette aux anarchies du despotisme, mise sur le trône aujourd'hui , détrônée demain. **(Temps, Definition Contextualisée)**

2. corpus.fr.utf8.t4t3t0.EcoleDoctorale3.these.Vigneron.txt.xml (Projet-theses, bibliosemantiqueETdefinition)
'La forme de la beauté est soit individuelle, - c'est-à-dire confinée à l'imitation d'un individu - soit une sélection de belles parties de nombreux individus et leur union en un tout, ce que nous appelons idéal, avec cependant la remarque qu'une chose peut être idéale sans être belle. **(Définition engagée)**

3. corpus.fr.utf8.t4t3t0.EcoleDoctorale3.these.Vigneron.txt.xml (Projet-theses, bibliosemantiqueETdefinition)
Il commence par nous dire que la grâce avait un sens différent chez les Grecs et il semble critiquer ce que ce terme est devenu dans la période que nous appelons le Rococo, il rejette ainsi cette " sorte d'affectation qui ne peut pas subsister en parfaite beauté sans l'embarrasser " et qui consiste " en certains gestes, actions et postures difficiles, non naturelles ou violentes ". **(Définition engagée)**

4. corpus.fr.utf8.t4t3t0.EcoleDoctorale3.these.Vigneron.txt.xml (Projet-theses, bibliosemantiqueETdefinition)
'Puisque la perfection n'a pas été accordée à l'humanité et ne peut être trouvée qu'en Dieu, et comme rien n'est compréhensible à notre nature excepté ce qui se trouve sous la conviction de nos sens, ainsi l'Omnipotence a trouvé bon d'imprimer un IDÉE visible de cette perfection qui est ce que nous appelons BEAUTÉ. **(Définition engagée)**

5. corpus.fr.utf8.t4t3t0.EcoleDoctorale3.these.Vigneron.txt.xml (Projet-theses, bibliosemantiqueETdefinition)
Gérard de Lairesse ne donne jamais de définition de la beauté mais la grâce et l'élégance en sont des parties inséparables. **(Opposition)**

Exemple : résultats

- Recherche des rencontres, mot clé « *Sarkozy* ».

Résultats

1. monde35corps.txt.xml (corpus1)
Les deux hommes s'étaient rencontrés en septembre 2006 lorsque Nicolas Sarkozy avait été reçu par le conseiller à la sécurité nationale de George Bush, Steve Hadley. **(Rencontre, Evénementielle, Individuelle, Réalisée)**

2. marseillaise_sifflee_laporte_propose_de_delocaliser_ce_.html.txt.xml (corpus4)
Tout match où "La Marseillaise" sera sifflée "sera immédiatement arrêté" et "les membres du gouvernement quitteront immédiatement l'enceinte sportive", a annoncé mercredi la ministre des Sports Roselyne Bachelot à l'issue d' une réunion à l'Elysée à laquelle Jean-Pierre Escalettes, le président de la Fédération française de football, avait été convoqué par Nicolas Sarkozy. **(Rencontre, Evénementielle, Réalisée)**

3. marseillaise_sifflee_tout_match_ou_lhymne_national_sera.html.txt.xml (corpus4)
Mme Bachelot s'exprimait à l'issue d' une réunion à l'Elysée à laquelle Jean-Pierre Escalettes, le président de la Fédération française de football (FFF), avait été convoqué par Nicolas Sarkozy. **(Rencontre, Evénementielle, Réalisée)**

4. monde35corps.txt.xml (corpus1)
Article paru dans l'édition du 05.08.07 L'Elysée et la Maison Blanche s'efforcent de déterminer les conditions d' une rencontre entre Nicolas Sarkozy et George Bush, à la faveur des vacances américaines du président français. **(Rencontre, Physique, Non réalisée)**

5. lib2corps.txt.xml (corpus1)
Lors d' une rencontre Merkel-Sarkozy-Kaczynski, une esquisse de compromis s'est dessinée et le président polonais était en contact avec son frère jumeau pour arrêter sa réponse. **(Rencontre, Evénementielle, Réalisée)**

Expressions régulières

Ressources linguistiques

- Listes de mots ou d'expressions (identifiables dans le texte) : on les appelle ***marqueurs linguistiques***.
- Règles :
  - SI tel marqueur et présent dans tel élément,
  - ALORS effectuer telle action.

- ***Expressions régulières*** :
  - C'est un formalisme, provenant de la *théorie des langages formels en mathématiques,* permettant de décrire certaines classes de chaînes de caractères, et donc certaines listes d'expressions linguistiques.

Expressions régulières (ER)

- Ce sont des chaînes de caractères construites en utilisant des caractères habituels (lettres de l'alphabet) et des opérateurs spécifiques.
- Chaque ER permet de décrire une classe de chaînes de caractères.
  - Cette classe s'appelle ***langage reconnu par l'expression régulière***.
- Si une classe de chaînes de caractères peut être reconnue par une ER (c'est-à-dire il existe une ER qui la reconnait), elle s'appelle ***un langage régulier***.

Expressions régulières : opérateurs

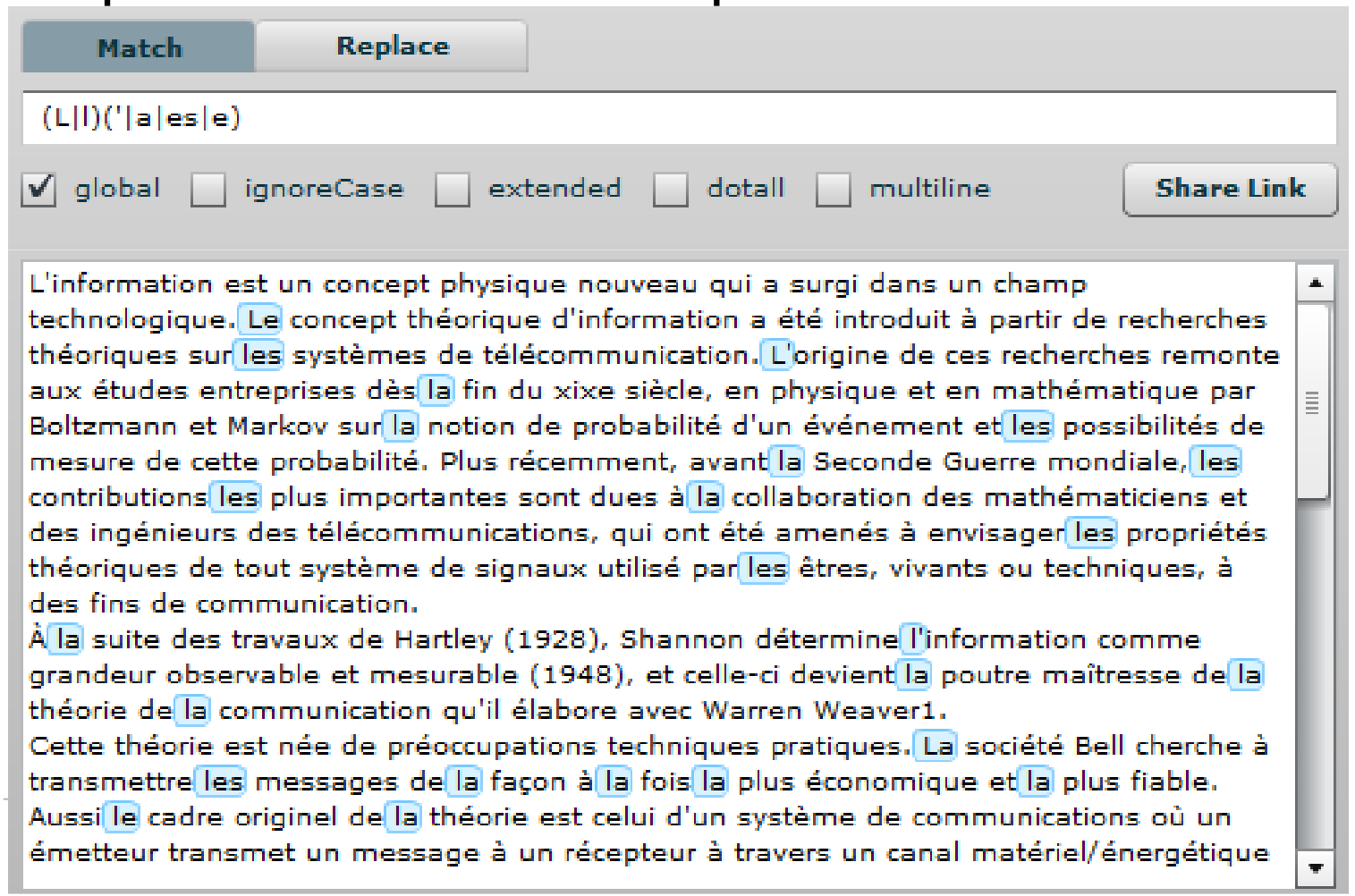
- Alternative **(…|…)** :
  - ***cardin(al | aux)*** reconnaît les mots : ***cardinal*** et ***cardinaux***.
- Etoile * : signifie que le caractère précédent est répété 0, un ou plusieurs fois :
  - ***ab**** reconnaît les chaînes : ***a, ab, abb, abbb, abbbb,*** …
  - ***a(bc)**** reconnaît les chaînes : ***a, abc, abcbc, abcbcbc,*** …
- **?** : signifie que le caractère précédent peut être présent ou pas :
  - ***boîtes?*** reconnaît les chaînes : ***boîte*** et ***boîtes***
  - ***ch(at|ien)s?*** reconnaît les chaînes : ***chat, chats, chien, chiens***

- ***[a-z]*** : une lettre de *a* à *z*.
- ***[0-9]*** : un chiffre de *0* à *9*.

Expressions régulières : caractères spéciaux

- ***\n*** : retour à la ligne
- ***\s*** : espace, tabulation ou retour à la ligne
- ***\w*** : n'importe quelle lettre ou chiffre ; équivaut à ***[0-9a-zA-Z]***
- ***\d*** : n'importe quel chiffre ; équivaut à ***[0-9].***

- Pour tester des ER : http://gskinner.com/RegExr/
- Cet outil permet de tester des expressions sur un texte.



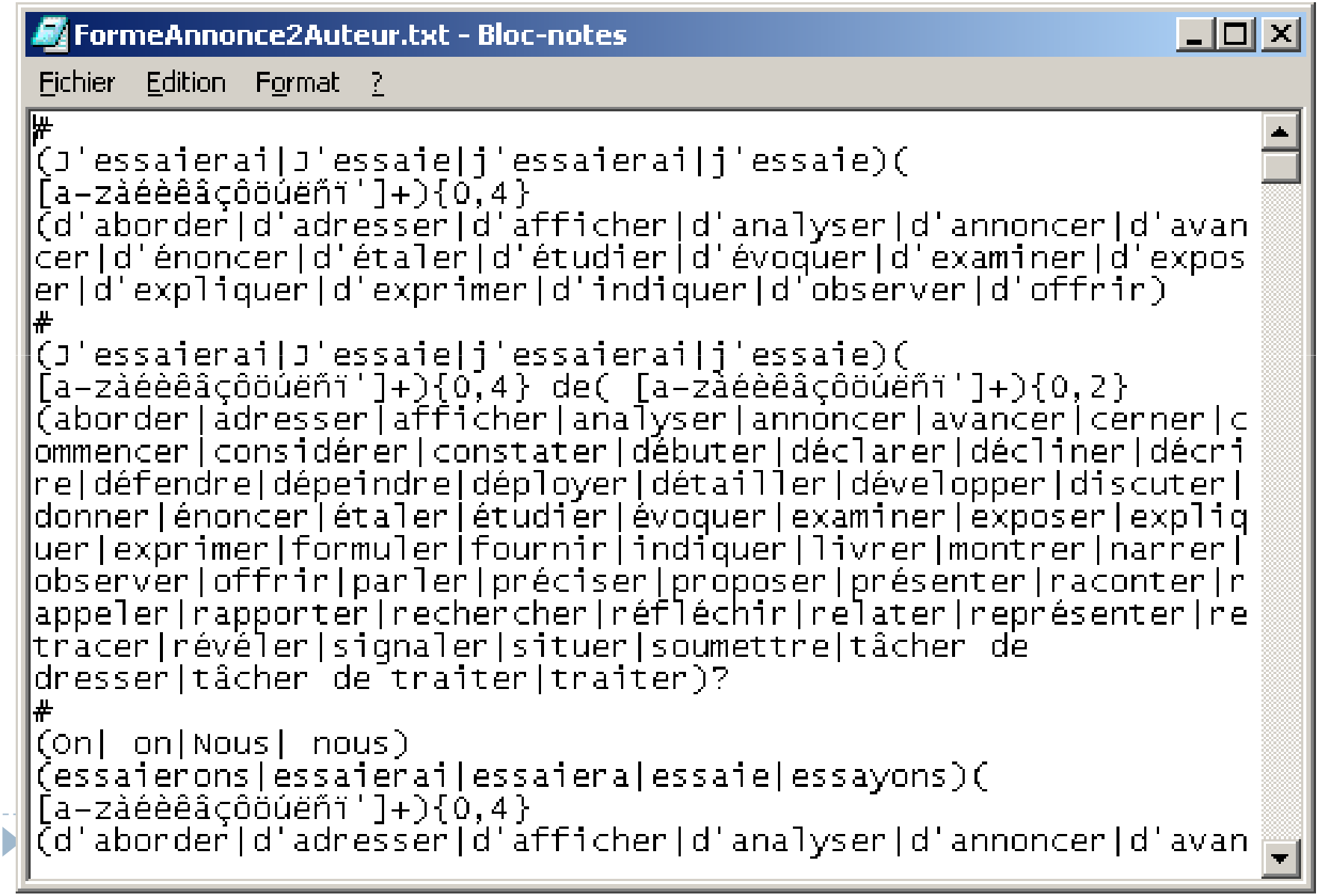
Exercice 1

- Quels sont les chaînes reconnues par les expressions régulières :
  - march(e|er|es|ons|ez|ont)
  - anima(l|ux) sauvages?
  - 19[0-9][0-9]
  - (je | nous) (peux|pouvons)? (\w)* suppos(er|e|ons) que

Exercice 2

- Construisez des expressions régulières qui reconnaissent :
  - l'article défini en français
  - les expressions :
    - *« je suis sûr(e) »*
    - *« je ne suis pas sûr(e)»*
    - *« je suis certain(e) »*
    - *« je ne suis pas certain(e) »*
    - *« je suis sûr(e) et certain(e) »*
  - la conjugaison du verbe *« finir »* en présent de l'indicatif

Exemple réel


FormeAnnonce2Auteur.txt - Bloc-notes

Fichier Edition Format ?

```
#
(J'essaierai|J'essaie|j'essaierai|j'essaie)(
[a-zàéèêâçôöúëñï']+){0,4}
(d'aborder|d'adresser|d'afficher|d'analyser|d'annoncer|d'avan
cer|d'énoncer|d'étaler|d'étudier|d'évoquer|d'examiner|d'expos
er|d'expliquer|d'exprimer|d'indiquer|d'observer|d'offrir)
#
(J'essaierai|J'essaie|j'essaierai|j'essaie)(
[a-zàéèêâçôöúëñï']+){0,4} de( [a-zàéèêâçôöúëñï']+){0,2}
(aborder|adresser|afficher|analyser|annoncer|avancer|cerner|c
ommencer|considérer|constater|débuter|déclarer|décliner|décri
re|défendre|dépeindre|déployer|détailler|développer|discuter|
donner|énoncer|étaler|étudier|évoquer|examiner|exposer|expliq
uer|exprimer|formuler|fournir|indiquer|livrer|montrer|narrer|
observer|offrir|parler|préciser|proposer|présenter|raconter|r
appeler|rapporter|rechercher|réfléchir|relater|représenter|re
tracer|révéler|signaler|situer|soumettre|tâcher de
dresser|tâcher de traiter|traiter)?
#
(On| on|Nous| nous)
(essaierons|essaierai|essaiera|essaie|essayons)(
[a-zàéèêâçôöúëñï']+){0,4}
(d'aborder|d'adresser|d'afficher|d'analyser|d'annoncer|d'avan
```

Exemple réel

- ERs pour reconnaître les renvoies bibliographiques dans un texte.

```
<marqueur no="401">(\[)[S][t][A-Z][a-z]+[A-Z][a-z]+[A-Z][a-z]+[0-9]{2}(\])</marqueur>
<marqueur no="402">\([A-Z][a-z]+(,( )?[A-Z][a-z]+){0,5}(,)?( )?(et ([A-Z][a-z]+))(,)?( )?[0-9]{4}\)</marqueur>
<marqueur no="403">\(voir par exemple [A-Z][a-z]+ et [A-Z][a-z]+, [0-9]{4}\)</marqueur>
<marqueur no="404">\(([A-Z][a-z]+(,( )?)?(([0-9]{4})(;|,)?( )?)?){0,5}(et( )?)[A-Z][a-z]+(,|;)( )?[0-9]{4}\)</marqueur>
<marqueur no="405">\(Adapté de ([A-Z][a-z]+(,( )?)?(([0-9]{4})(;|,)?( )?)?){0,5}(et( )?)[A-Z][a-z]+(,|;)( )?[0-9]{4}\)</marqueur>
<marqueur no="406">\(([A-Z][\'])?[A-Z][a-z]+( )et( )([A-Z][\'])?[A-Z][a-z]+,( )?[0-9]{4}\)</marqueur>
<marqueur no="407">\(([A-Z][a-z]+(, )?(, et )?){0,5}[A-Z][a-z]+(, et )?(, [0-9]{4}), [0-9]{4}\)</marqueur>
<marqueur no="408">\(Adapté de ([A-Z][a-z]+(,( )?)?){0,5}( et)?( )?([A-Z][a-z]+,( )?[0-9]{4})\)</marqueur>
<marqueur no="409">\(D'après ([A-Z][a-z]+(,( )?)?){0,5}( et)?( )?([A-Z][a-z]+,( )?[0-9]{4})\)</marqueur>
<marqueur no="410">\(([A-Z][\'])?[A-Z][a-z]+(( )et( ))?(, )?){1,6}(et ([A-Z][\'])?[A-Z][a-z]+)(, )[0-9]{4}\)</marqueur>
```

Limites

- Les ERs permettent de reconnaître des classes très limitées de chaînes de caractères.
- La langue naturelle est beaucoup plus complexe que n'importe quelle classe reconnaissable par des ERs (suite d'une théorème de Chomsky et la classification des langages formels).

- Utiliser d'autres outils plus puissants ?

Format XML

Texte annoté : XML

- L'information rajoutée est stockée dans un format spécifique, souvent XML (Extended Markup Language).
- C'est un format qui permet de définir différents éléments textuels qui font partie d'une structure arborescente. Chaque élément a un nom, un contenu textuel et des attributs.

- Il permet alors à délimiter les différents éléments textuels et d'en rajouter des informations supplémentaires.

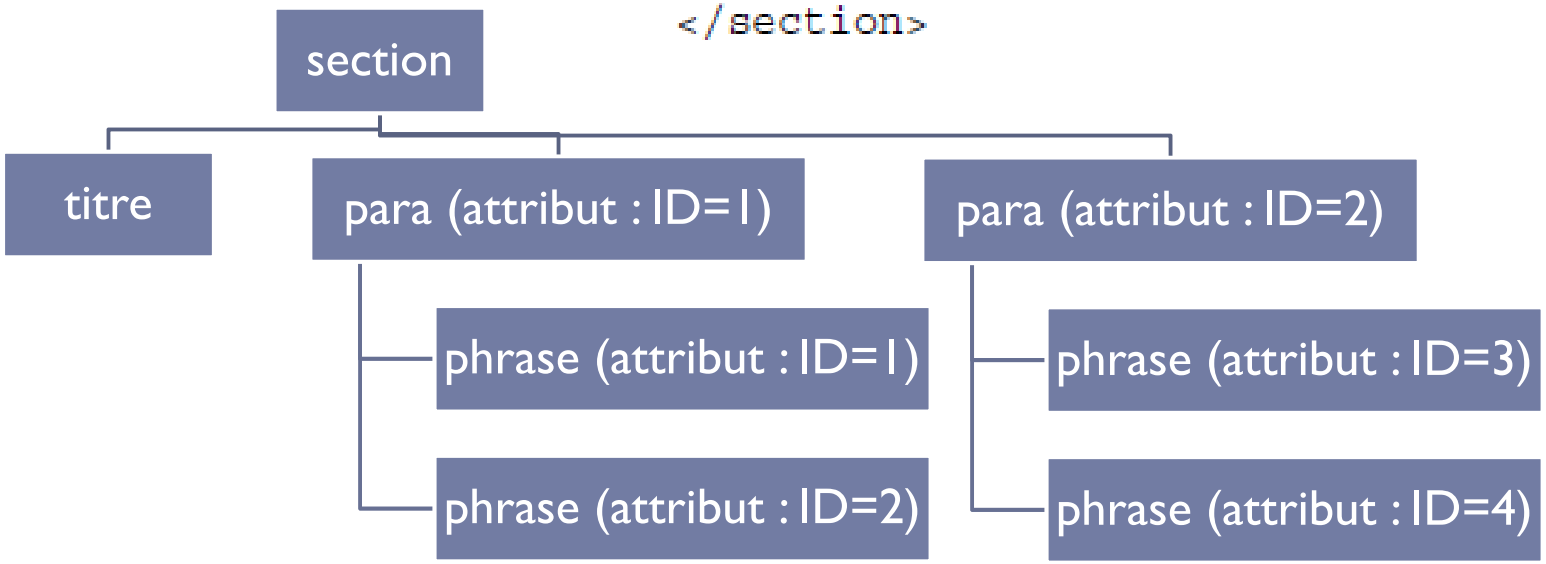
XML

- Chaque élément :
  - commence par une balise ouvrante : par ex. ***<phrase>***
  - se termine par une balise fermante : par ex. ***</phrase>***
- Le contenu textuel et les sous-éléments se trouvent entre ces deux balises.
- L'élément peut avoir des attributs, qui seront mentionnés dans la balise ouvrante : par ex.

```
<phrase attr1="…" attr2="…" attr3="…" >
        <mot>L'oiseau</mot>
        <mot>est</mot>
        <mot>sur</mot>
        <mot>la</mot>
        <mot>branche </mot> .
</phrase>
```

Exemple XML

```
<section ID=1>
<titre>Title</titre>
<para ID=1>
<phrase ID=1>First sentence.</phrase>
<phrase ID=2>Second sentence. </phrase>
</para>
<para ID=2>
<phrase ID=3>Third sentence. </phrase>
<phrase ID=4>Fourth sentence. </phrase>
</para>
</section>
```



Exemple réel d'une phrase annotée

```
<phrase id="2" idTotal="37" annotation="similitude"
              regle="..\ressources\bibliosemantique\regles\regles_similitude.xml"
              listeIndicateurs="..\ressources\bibliosemantique\indicateurs\indicateurs\marqueurs.xml">
        <avantIndicateur>
              <avantIndice>En effet,</avantIndice>
              <indice>nous souscrivons tout à fait à l'analyse de</indice>
              <apresIndice>Carol, Briggs et Bange (</apresIndice>
        </avantIndicateur>
        <indicateur>[CarolEtAl04] : 10</indicateur>
        <apresIndicateur>).</apresIndicateur>
</phrase>
```

XML et HTML

- XML est un format textuel.
- Le format HTML des pages web est un cas particulier de XML : dans le format HTML les noms des éléments sont définis de façon à permettre aux navigateurs web d'interpréter les éléments afin de positionner leur contenu sur la page web.
- Pour afficher le code source d'une page : *clic droit > Afficher le code source*.
- Exemple :

```
<div class="post" id="cours11">

<h2>Cours 11 : TAL : Annotation sémantique et ressources linguistiques</h2>

    <date> 12/12/2011 - 15/12/2011 </date>

    <ul>
    <li><a href="Cours11.pdf">Cours 11</a></li>
    <li><a href="Cours11_version_imprimable.pdf">Cours 11 (version imprimable)</a></li>
    </ul>

</div>
```