Colloque international et école d'été

Albi, 10-14 juillet 2006

CORPUS EN LETTRES ET SCIENCES SOCIALES :
DES DOCUMENTS NUMÉRIQUES
À L'INTERPRÉTATION

Rey Christophe, Zaoui Corinne
Université de Provence,
Equipe DELIC
Christophe.Rey@up.univ-aix.fr
zaoui@up.univ-aix.fr

La résurrection du dictionnaire ancien par la déconstruction positive de l'informatique

Les solutions existantes

Deux grandes solutions

Le balisage Minimal

Le balisage Analytique

Le balisage minimal

TOUCHER LE MOINS POSSIBLE AU TEXTE

Wooldridge
Leroy-Turcan
Caron / Dagenais

Caractéristiques

- **Minimalisation** de l'analyse microstructurelle
- Jalons sur les **champs informationnels les plus facilement identifiables**
- Balisage d'**informations typographiques**

(Edition, pages, colonne, alinéas, etc.)

- Listes de **mots-clés métalinguistiques**

→

- ***Thrésor de la langue françoise*** de Nicot (1606)
- ***Dictionnaire de l'Académie Française***
- Etc.

Exemple de balisage Minimal

[...] TIMBRE. s. m. Sorte de cloche ronde qui n'a point de battant en dedans, & qui est frappée en dehors par un marteau. *Le timbre d'une horloge. timbre d'un reveille-matin. le timbre de cette horloge est tres-bon*.
[...]
Timbrer. v.a. Terme de blason, Accompagner d'un timbre. *Timbrer une armoirie*.
Timbrer. v.a. Terme de Pratique, Escrire en haut d'un Acte, la nature de cet acte, sa date & le sommaire de ce qu'il contient. *Timbrer des pieces*.
On dit aussi, *Timbrer du papier, timbrer du parchemin*, pour dire, Imprimer la marque du Roy sur du papier, sur du parchemin, pour faire qu'il puisse servir aux actes de Justice.

Localisation de la lexie dans la structure dictionnairique

Information typographique sur la forme

<page n="563"><col n="1">[...] <lc>TIMBRE</lc>. s. m. Sorte de cloche ronde qui n'a point de battant en dedans, & qui est frappée en dehors par un marteau. Le timbre d'une horloge. timbre d'un reveille-matin. le timbre de cette horloge est tres-bon.
[...]
<sc>Timbrer</sc>. v.a. Terme de blason, Accompagner d'un timbre. Timbrer une armoirie.
<sc>Timbrer</sc>. v.a. Terme de Pratique, Escrire en haut d'un Acte, la nature de cet acte, sa date & le sommaire de ce qu'il contient. Timbrer des pieces.
On dit aussi, Timbrer du papier, timbrer du parchemin, pour dire, Imprimer la marque du Roy sur du papier, sur du parchemin, pour faire qu'il puisse servir aux actes de Justice.

Restitution de la mise en page

Aspect italique de certaines informations

Le balisage Analytique

TOUT BALISER

Wionet / Tutin

Caractéristiques

- Balisage **SGML**
- Etude **approfondie** de la microstructure
- Jalons **sur l'ensemble des champs informationnels** de l'article
- Balisage d'**informations typographiques**

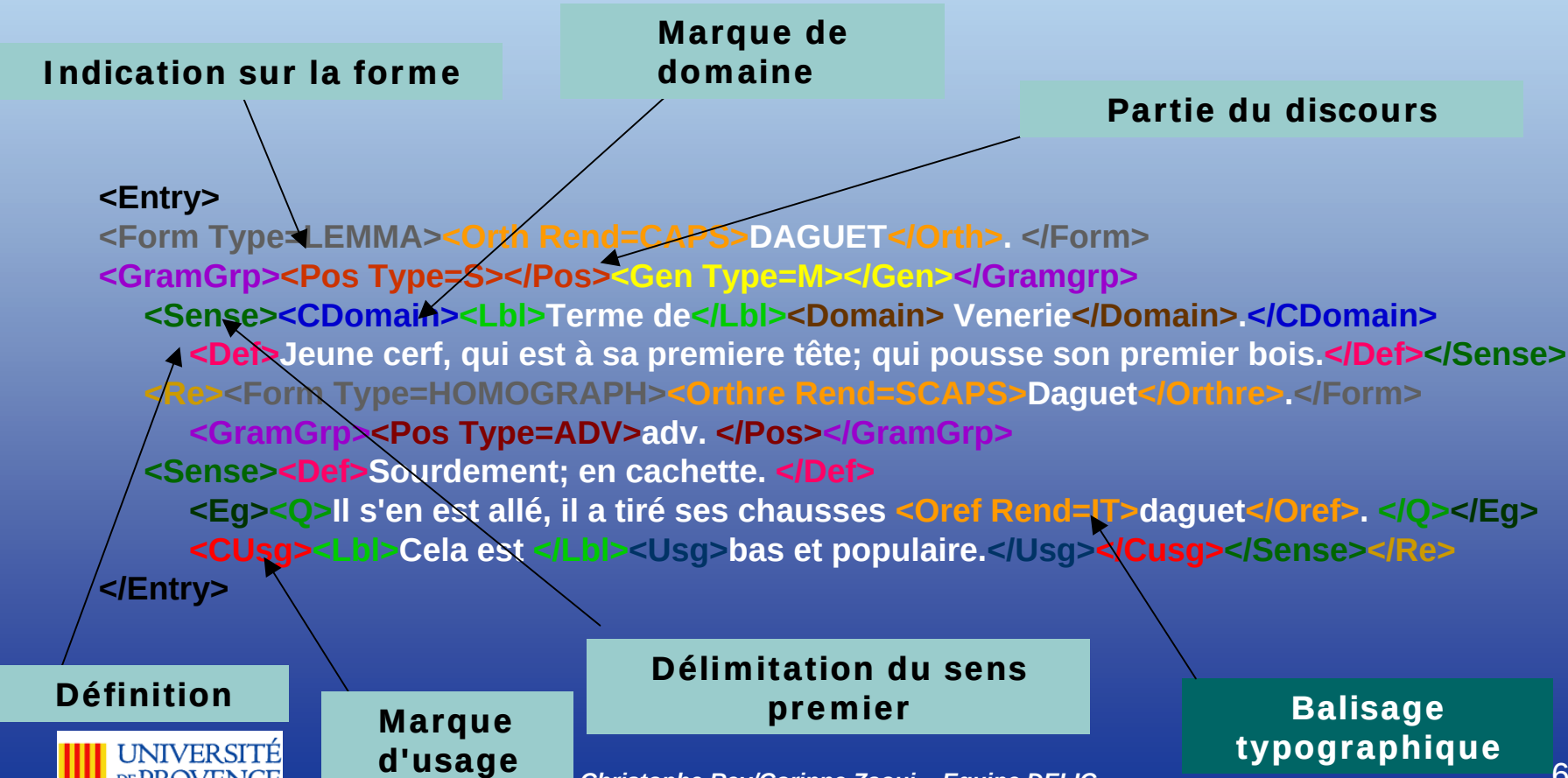
(Petites capitales, italiques, etc.)

→

- ***Dictionnaire Etymologique ou Origine de la Langue Françoise*** de Gilles Ménage (1694),
- ***Dictionnaire Universel*** **de Furetière revu par Basnage de Bauval** (1702).

Exemple de balisage Analytique

DAGUET. Terme de Venerie. Jeune cerf, qui est à sa premiere tête; qui pousse son premier bois. Daguet. adv. Sourdement; en cachette. Il s'en est allé, il a tiré ses chausses *daguet*. Cela est bas et populaire.



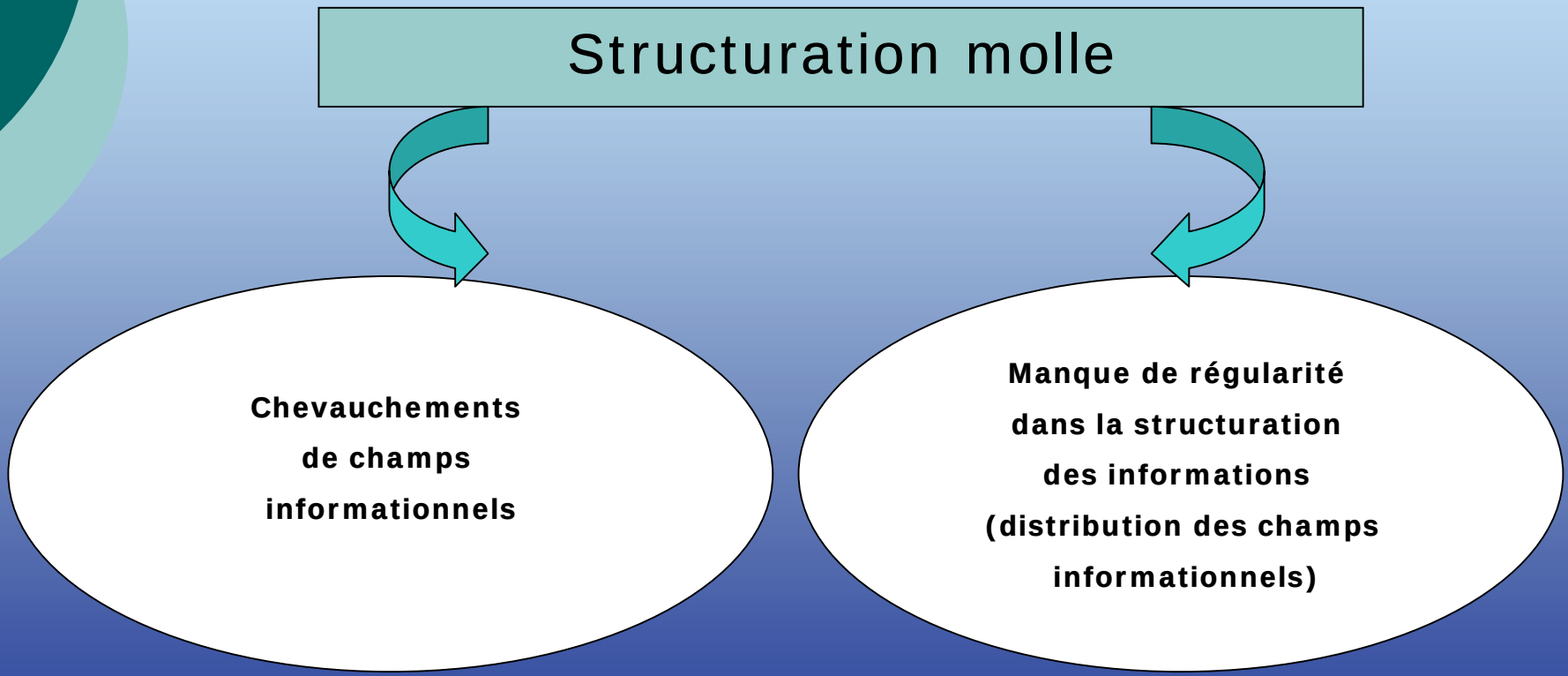
Un corpus spécifique

Le dictionnaire *Grammaire & Littérature* (1782-1784-1786)

GRAMMAIRE	LITTÉRATURE
 Nicolas Beauzée (1717-1789)	 Jean-François Marmontel (1723-1799)

Corpus composé de 236 lexies

Quelques problèmes spécifiques



Difficultés de découpage des champs informationnels

K, s.m. *Grammaire*. Si l'on confond à l'ordinaire l'*i* voyelle & l'*i* consonne, *K* est la dixième lettre & la septième consonne de notre alphabet ; mais si l'on distingue, comme je l'ai fait, la voyelle *I* & la consonne *J*, il faut dire que *K* est la onzième lettre & la huitième consonne de notre alphabet ; & c'est d'après cette hypothèse très-raisonnable, que désormais je coterai les autres lettres.

K,	**Forme**
s.m.	**Information grammaticale**
***Grammaire*.**	**Marque de domaine**
Si l'on confond à l'ordinaire l'*i* voyelle & l'*i* consonne, *K* est la dixième lettre & la septième consonne de notre alphabet ; mais si l'on distingue, comme je l'ai fait, la voyelle *I* & la consonne *J*, il faut dire que *K* est la onzième lettre & la huitième consonne de notre alphabet ; & c'est d'après cette hypothèse très-raisonnable, que désormais je coterai les autres lettres.	**Définition**
	Développement encyclopédique
	Définition

Déconstruire positivement le texte ancien grâce au balisage souple ou flottant

Caractéristiques du balisage souple ou flottant

- **Balisage XML**
- **Balisage logique**
- **Balisage physique dissocié**
- **S'affranchit d'une structuration stricte de l'ensemble des données textuelles**

Le balisage "souple" ou "flottant" (1)

(N.) DIGAMMA, s. m. Double Gamma. On a donné anciennement ce nom à la lettre F, qui paroit en effet composée de deux Gamma placés verticalement l'un sur l'autre. Voyez F. (M.BEAUZÉE.)

```
<ARTICLE>
    <STATUT TYPE="NOUVEAU"/>
        <ENTREE TYPE="EP">
            <FORME>(N. ) DIGAMMA</FORME>,
            <INFORMATION_GRAMMATICALE TYPE="SUBSTANTIF MASCULIN">
                <PARTIE_DU_DISCOURS TYPE="SUBSTANTIF">s.
                </PARTIE_DU_DISCOURS>
                <GENRE TYPE="MASCULIN">m.
                </GENRE>
            </INFORMATION_GRAMMATICALE>
        </ENTREE>
        <CORPS>
            <DEFINITION>Double Gamma.</DEFINITION>
            <DISCOURS_ENCYCLOPEDIQUE>On a donné anciennement ce nom à la lettre F,
            qui paroit en effet composée de deux Gamma placés verticalement l'un sur l'autre.
            <REFERENCE TYPE="VEDETTE">Voyez F</REFERENCE>.
            </DISCOURS_ENCYCLOPEDIQUE>
            <SIGNATURE TYPE="BEAUZEE">(M.BEAUZÉE.)</SIGNATURE>
        </CORPS>
</ARTICLE>
```

Le balisage "souple" ou "flottant" (2)

(N.) GUTTURAL, E, adj. Appartenant à la gorge ou au gosier. Vaisseau guttural. Glande gutturale. Articulations, Consonnes gutturales. Ce mot, tiré immédiatement du latin Gutturalis, qui a le même sens, vient du nom Guttur (Gorge, Gosier). Les articulations gutturales sont celles qui font retentir l'explosion de la voix dans la région du gosier. Il y en a deux bien sensibles dans le françois, G & Q ; telles qu'on les entend dans les mots Gale, Cale; vaguer, vaquer; &c. (M.BEAUZÉE.)

```
<ARTICLE>
    <STATUT TYPE="TRES DIFFERENT"/>
        <ENTREE TYPE="EP">
            <FORME>(N. ) GUTTURAL, E</FORME>,
            <INFORMATION_GRAMMATICALE TYPE="ADJECTIF">
                <PARTIE_DU_DISCOURS TYPE="ADJECTIF">adj.
                </PARTIE_DU_DISCOURS>
            </INFORMATION_GRAMMATICALE>
    </ENTREE>
    <CORPS>
        <DEFINITION>Appartenant à la gorge ou au gosier.</DEFINITION>
        <CONTEXTUALISATION>Vaisseau guttural. Glande gutturale. Articulations, Consonnes
        gutturales.</CONTEXTUALISATION><ETYMOLOGIE>Ce mot, tiré immédiatement du
        latin <LANGUE TYPE="LATIN">Gutturalis </LANGUE>, qui a le même sens, vient du
        nom <LANGUE TYPE="LATIN">Guttur </LANGUE> (Gorge, Gosier).
        </ETYMOLOGIE><DISCOURS_ENCYCLOPEDIQUE>Les articulations gutturales sont
        celles qui font retentir l'explosion de la voix dans la région du gosier. Il y en a deux bien
        sensibles dans le françois, G & Q ; telles qu'on les entend dans les mots
        <EXTRA TYPE="PAIRE MINIMALE">Gale, Cale </EXTRA>;
        <EXTRA TYPE="PAIRE MINIMALE">vaguer, vaquer </EXTRA>;
        &c.</DISCOURS_ENCYCLOPEDIQUE>
        <SIGNATURE TYPE="BEAUZEE">(M.BEAUZÉE.)</SIGNATURE>
    </CORPS>
</ARTICLE>
```

Le balisage "souple" ou "flottant" (3)

(N.) MOBILE, adj. Susceptible de mouvement. Les hébraïsants qui suivent la méthode massorétique nomment lettres mobiles, celles qui se prononcent toujours ; parce qu'elles sont, dit l'abbé Ladvocat Gramm. hébr. PAG. 7.), comme mises en mouvement par les organes de la voix. Toutes les lettres hébraïques sont mobiles, à la réserve de quatre, que les massorètes nomment, par opposition, Quiescentes. Voy. ce mot. (M.BEAUZÉE.)

```
<ARTICLE>
        <ENTREE TYPE="EP">
            <FORME>(N.) MOBILE</FORME>,
            <INFORMATION_GRAMMATICALE TYPE="ADJECTIF">
                <PARTIE_DU_DISCOURS TYPE="ADJECTIF">adj.
                </PARTIE_DU_DISCOURS>
            </INFORMATION_GRAMMATICALE>
        </ENTREE>
        <CORPS>
            <DEFINITION>Susceptible de mouvement.</DEFINITION>
            <DISCOURS_ENCYCLOPEDIQUE>Les hébraïsants qui suivent la méthode
            massorétique nomment lettres mobiles, celles qui se prononcent toujours ; parce
            qu'elles sont, dit l'<REFERENCE TYPE="PERSONNE">abbé Ladvocat
            </REFERENCE>(<REFERENCE TYPE="OUVRAGE">Gramm. hébr. PAG.
            7.</REFERENCE>), comme mises en mouvement par les organes de la voix. Toutes
            les lettres hébraïques sont mobiles, à la réserve de quatre, que les massorètes
            nomment, par opposition, <REFERENCE TYPE="VEDETTE">Quiescentes. Voy.
            ce mot.</REFERENCE>
            </DISCOURS_ENCYCLOPEDIQUE>
            <SIGNATURE TYPE="BEAUZEE">(M.BEAUZÉE.)</SIGNATURE>
        </CORPS>
</ARTICLE>
```

Exploitation du balisage souple ou flottant : le logiciel CorpXML

Présentation de CorpXML

Un logiciel libre

- C++ en association avec les librairies QT
- API DOM (Document Object Model)

http://www.up.univ-mrs.fr/delic/perso/rey/methodique/index.htm

Références bibliographiques

-CARON, P., DAGENAIS, L., GONFROY, G. (1992). "Le programme d'informatisation du Dictionaire critique de la langue française de l'abbé Jean-François Féraud (1787)", Historical Dictionary Databases (éd. T.R. Wooldridge). *CCH Working Papers*, 2: pp. 87-103.

- DENDIEN, J, PIERREL, J-M. (2003). "Le trésor de la Langue Française informatisé : un exemple d'informatisation d'un dictionnaire de langue de référence", *TAL*. Volume 44 - n°2/2003, 28 p.

- REY, C. (2004). Analyse et informatisation des articles traitant de l'étude des sons dans le dictionnaire Grammaire & Littérature de Nicolas Beauzée et Jean-François Marmontel, issu de l'Encyclopédie Méthodique. Thèse de doctorat. Aix-en-Provence.

- REY, C., ZAOUI, C. (2004). Le balisage XML "ciblé" : une nouvelle approche dans l'informatisation des corpus. Actes de la conférence internationale sur la Fouille de Texte (CIFT'04), dans le cadre de la semaine du Document Numérique, La Rochelle, 22-24 juin 2004, pp. 121-133.

- VÉRONIS, J., & IDE, N. (1996). Encodage des dictionnaires électroniques: problèmes et propositions de la TEI. In D. Piotrowsky (Ed.), Lexicographie et informatique - Autour de l'informatisation du Trésor de la Langue Française. Actes du Colloque International de Nancy (29, 30, 31 mai 1995), pp. 239-261, Paris: Didier Erudition.

- WIONET C., TUTIN A. (2001). Pour informatiser le Dictionnaire universel de Basnage (1702) et de Trévoux (1704). Approche théorique et pratique. Honoré Champion.

- WIONET C., TUTIN A. (1998). "Informatisation du Dictionnaire Universel de Furetière revu par Basnage de Bauval (1702) : premier bilan", *Actes du colloque-atelier international DictA1998 organisé par le Groupe d'Études sur l'Histoire de la Langue Française (GEHLF) et la Société Internationale d'Études Historiques et Linguistiques des Dictionnaires Anciens (SIEHLDA)*, Université de Limoges, 19-20 novembre 1998.

- WOOLDRIDGE, T.R. (1998). "L'informatisation du Dictionnaire de l'Académie française (DAF)", dans *Actes du colloque-atelier international DictA1998 organisé par le Groupe d'Études sur l'Histoire de la Langue Française (GEHLF) et la Société Internationale d'Études Historiques et Linguistiques des Dictionnaires Anciens (SIEHLDA)*, Université de Limoges, 19-20 novembre 1998.

- WOOLDRIDGE, T.R., LEROY-TURCAN I. (1996). "Les mots-clefs métalinguistiques comme outil d'interrogation structurante des dictionnaires anciens", Lexicomatique et dictionnairiques (éd. A. Clas, P. Thoiron & H. Béjoint), Beyrouth: FMA & Montréal: AUPELF-UREF, pp. 307-16.

- WOOLDRIDGE, T.R. (1994). "Projet d'informatisation du Dictionnaire de l'Académie (1694-1935)", Actes du Colloque international Le Dictionnaire de l'Académie française et la lexicographie institutionnelle européenne, Institut de France, novembre 1994; (ed. B. Quemada & J. Pruvost), Paris, Champion: 309-20.

- WORLD WIDE WEB CONSORTIUM. Extensible Markup Language (XML) : http://www.w3.org/XML/.