TEXTCOOP

Classification automatique en genres :

Types de traits utilisés pour représenter les textes. *Relevé bibliographique*.

Maria ZIMINA
LIPN (Paris 13)

Plan

- Préambule typologique :
  - Caractérisations *a priori* / *a posteriori*
  - Notions de *genre*, *type*, *fonction discursive*
- Traits utilisés pour la typologie en genres
  - Niveaux de traits
  - Exemples
- Problèmes de tels ensembles de traits
  - Hétérogénéité textuelle
  - Décomptes à plat (textes = « sacs d'indices » ??)

Préambule typologique 1/4

- Caractérisation *a priori*
  - caractéristiques externes des textes (fonctionnelles, situationnelles)
- Caractérisation *a posteriori*
  - caractéristiques internes des textes (caractérisation linguistique)

☛ D'où une distinction entre :

genres ou *registres* (définis sur des critères situationnels)

et *types de textes* (définis sur des critères linguistiques internes)

Marie-Paule Péry-Woodley (2001) « Mode d'organisation et de signalisation dans les textes procéduraux ». *Langages n°141*.

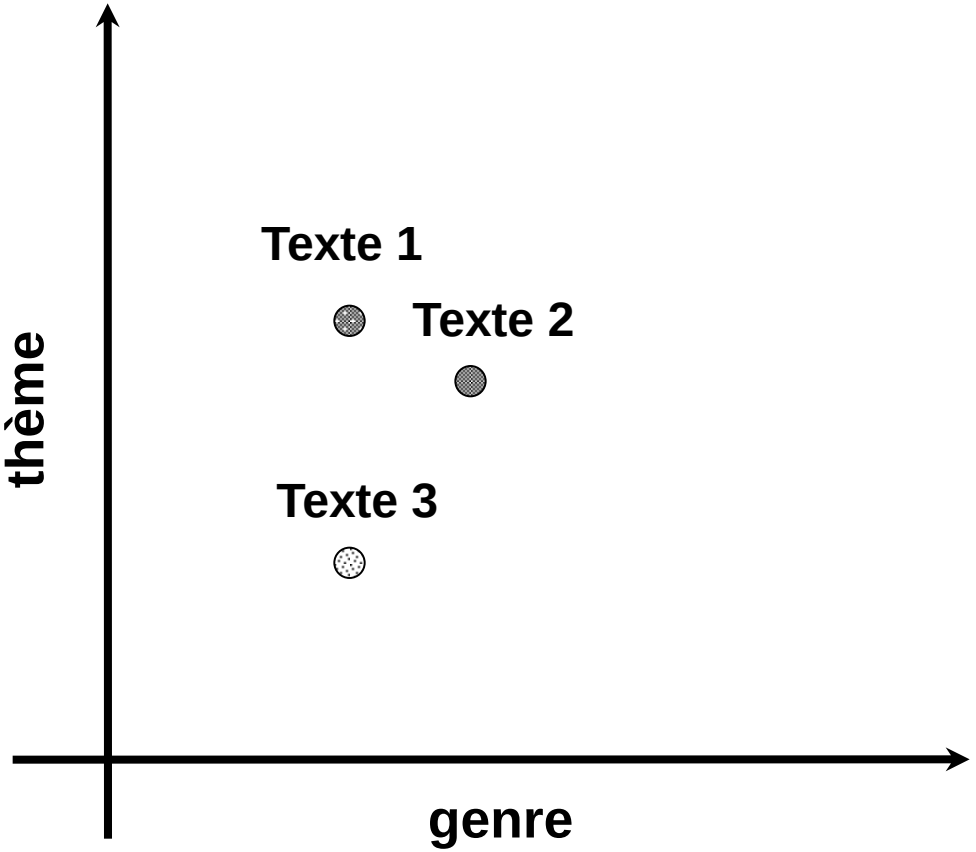
Préambule typologique 2/4

- La *typologie en genres* est évolutive :
  - Elle est constamment appelée à refléter l'évolution des pratiques discursives (exemple : apparition de l'Internet…)

- La *typologie inductive* est fondée sur des traits internes :
  - Justification des traits ? Interprétation des regroupements (types de textes) observés ?

Préambule typologique 3/4

- Classification thématique *vs*. classification en genres



Préambule typologique 4/4

- Classification thématique
  - **Domaine 'droit' :** code civil, jurisprudences, comptes-rendus d'audiences, etc.

- Classification en genres
  - Éclairage sémantique global
  - Problème de repérage : sur quels fonctionnements linguistiques ?

Typologie en genres : traits utilisés pour le traitement automatique

- Niveau des traits utilisés :
  - niveau caractères / formes graphiques
  - niveau lexical
  - niveau morpho-syntaxique (POS)
  - niveau syntaxique (analyse de dépendances syntaxiques)
  - niveau sémantique
  - niveau typo-dispositionnel
  - caractéristiques générales du document

Niveau caractères/formes graphiques

- sigles commerciaux ou monétaires **($, £, ©, ®, ™ …)**
- dates (prise en compte des formats)
- acronymes **(FAQ, POS, …)**
- ponctuation **(.,:;!?« »" …)**

- Exemples de ressources :
  - listes prédéfinies de sigles : http://acronymes.info

Niveau lexical / formes graphiques

- ensembles de mots prédéfinis :
  - les plus fréquents dans une langue…
  - entités nommées, …
  - termes (degré de spécialisation)
  - mots mal orthographiés, …

➔ Exemples de ressources :
  - ➔ dictionnaires de langue générale
  - ➔ dictionnaires de langue de spécialité

Exemple d'un jeu de traits (niveau caractères / niveau lexical) :

	Feature type	Feature set A
(10)	Closed word sets	avg. word frequency class
(11)		avg. # of currency symbols
(12)		avg. # of help symbols
(13)		avg. # of shop symbols
(14)		avg. # of date symbols
(15)		avg. # of first names
(16)		avg. # of surnames
(17)		avg. # of words that do not appear in Webster's dictionary
(18)	Text statistics	avg. # of question marks
(19)		avg. # of letters
(20)		avg. # of digits
(21)		avg. # of dots
(22)		avg. # of semicolons
(23)		avg. # of colons
(24)		avg. # of commas
(25)		avg. # of exclamation marks

Sven Meyer zu Eissen, Benno Stein (2004) « Genre Classification of Web Pages », In *Proceedings of KI-2004*.

Niveau morpho-syntaxique

- Étiquetage morpho-syntaxique (*Part of Speech Tagging - POS*)
  - temps des verbes
  - mode (infinitif, passif,...)
  - prépositions
  - pronoms : personne
  - déterminants
  - adverbes…

- Exemples de ressources (outils):
  - ***TreeTagger*** (Étiqueteur) : http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/
  - ***Flemm*** (Analyseur Flexionnel pour des corpus étiquetés) : http://www.univ-nancy2.fr/pers/namer/Telecharger_Flemm.html

TreeTagger : jeu d'étiquettes pour le Français

1.	ABR	ABREVIATION
2.	ADJ	ADJECTIVE
3.	ADV	ADVERB
4.	DET:art	ARTICLE
5.	DET:pos	POSSESSIVE PRONOUN (ma, ta, ...)
6.	INT	INTERJECTION
7.	KON	CONJUNCTION
8.	NAM	PROPER NAME
9.	NOM	NOUN
10.	NUM	NUMERAL
11.	PRO	PRONOUN
12.	PRO:dem	DEMONSTRATIVE PRONOUN
13.	PRO:ind	INDEFINITE PRONOUN
14.	PRO:per	PERSONAL PRONOUN
15.	PRO:pos	POSSESSIVE PRONOUN (mien, tien, ...)
16.	PRO:rel	RELATIVE PRONOUN
17.	PRP	PREPOSITION
18.	PRP:det	PREPOSITION PLUS ARTICLE (au,du,aux,des)
19.	PUN	PUNCTUATION
20.	PUN:cit	PUNCTUATION CITATION
21.	SENT	SENTENCE TAG
22.	SYM	SYMBOL
23.	VER:cond	VERB CONDITIONAL
24.	VER:futu	VERB FUTUR
25.	VER:impe	VERB IMPERATIVE
26.	VER:impf	VERB IMPERFECT
27.	VER:infi	VERB INFINITIVE
28.	VER:pper	VERB PAST PARTICIPLE
29.	VER:ppre	VERB PRESENT PARTICIPLE
30.	VER:pres	VERB PRESENT
31.	VER:simp	VERB SIMPLE PAST
32.	VER:subi	VERB SUBJUNCTIVE IMPERFECT
33.	VER:subp	VERB SUBJUNCTIVE PRESENT

Sorties de TreeTagger (corpus TextCoop) 1/2

Le	**DET:ART**	**le**
mortier	**NOM**	**mortier**
est	**VER:pres**	**être**
à	**PRP**	**à**
la	**DET:ART**	**le**
base	**NOM**	**base**
de	**PRP**	**de**
tous	**PRO:IND**	**tout**
les	**DET:ART**	**le**
travaux	**NOM**	**travail**
de	**PRP**	**de**
maçonnerie	**NOM**	**maçonnerie**
que	**KON**	**que**
ce	**PRO:DEM**	**ce**
soit	**VER:subp**	**être**
pour	**PRP**	**pour**
construire	**VER:infi**	**construire**
,	**PUN**	**,**
pour	**PRP**	**pour**
enduire	**VER:infi**	**enduire**
ou	**KON**	**ou**
pour	**PRP**	**pour**
réparer	**VER:infi**	**réparer**
.	**SENT**	**.**

Sorties de TreeTagger (corpus TextCoop) 2/2

```
<?xml version="1.0" encoding="iso-8859-1" standalone="no"?>
<?xml-stylesheet type="text/xsl" href="treetagger.xsl"?>
<document>
<article>
<html>
        <element>
                <data type="type">NOM</data>
                <data type="lemma">mortier</data>
                <data type="string">mortier</data>
        </element>

        <element>
                <data type="type">VER:pres</data>
                <data type="lemma">être</data>
                <data type="string">est</data>
        </element>
/…/
```

Travaux ‘Plurital’ *2005-06* : transformation de sorties TreeTagger -> XML

Flemm : analyseur flexionnel

- calcule le lemme de chaque mot fléchi en fonction de l'étiquette fournie en entrée par :
  - ***TreeTagger***
  - ***Brill***
- fournit les traits flexionnels principaux :
  - genre et nombre pour les adjectifs, déterminants, participes ;
  - nombre pour les noms ;
  - genre, nombre, personne et cas pour les pronoms ;
  - nombre, personne, temps, mode et groupe de conjugaison pour les verbes.

Sorties de Flemm 1/2

Le	DET(ART):Da3ms---	le
mortier	NOM:Nc-s--	mortier
est	VER(pres):Vmip3s--3	être
à	PRP	à
la	DET(ART):Da3fs---	le
base	NOM:Nc-s--	base
de	PRP	de
tous	PRO(IND):Pi3mp--	tout
les	DET(ART):Da3-p---	le
travaux	NOM:Ncmp--	travail
de	PRP	de
maçonnerie	NOM:Nc-s--	maçonnerie
que	KON	que
ce	PRO(DEM):Pd3msn-	ce
soit	VER(subp):Vmip3s--3	être || soit
	VER(subp):Vmis3s--3	être
pour	PRP	pour
construire	VER(infi):Vmn------	construire
,	PUN	,
pour	PRP	pour
enduire	VER(infi):Vmn------	enduire
ou	KON	ou
pour	KON	pour
réparer	VER(infi):Vmn------	réparer
.	SENT	.

Sorties de Flemm 2/2

```
<?xml version='1.0' encoding='ISO-8859-1'?>

<FlemmResult>
        <InflectedForm>est</InflectedForm>
        <Category original-tagger='VER:pres'>VER(pres)</Category>
        <Analyses> <!-- est        VER(pres):Vmip3s--3      être -->
                <Analyse>
                        <Lemme>être</Lemme>
                        <Features>
                                <Feature name='catmultext' value='V'/>
                                <Feature name='type' value='m'/>
                                <Feature name='mood' value='i'/>
                                <Feature name='tense' value='p'/>
                                <Feature name='pers' value='3'/>
                                <Feature name='gend' value='-'/>
                                <Feature name='nb' value='s'/>
                                <Feature name='clitic' value='-'/>
                                <Feature name='vclass' value='3'/>
                        </Features>
                </Analyse>
        </Analyses>
</FlemmResult>
```

N-grams de POS

- Niveau morpho-syntaxique portant sur des n-grams (POS *n-grams*)
  - souvent des trigrams (M. Santini, University of Brighton, M. Gamon, Microsoft Research, etc.) :
  - Exemple (pour l'anglais) :
    - NOUN CONJ NOUN / NOUN PREP ADJ
    - Fréquence d'apparition en corpus (entre 70% et 40%)

- Exemples de ressources (outils -> à voir):
  - Ngram Statistics Package (NSP) de Ted Pedersen : http://www.d.umn.edu/~tpederse/nsp.html

    Identification de *n-gram* par tests d'association (*Fisher*, *Dice, chi-2, log likelihood ratio…*)

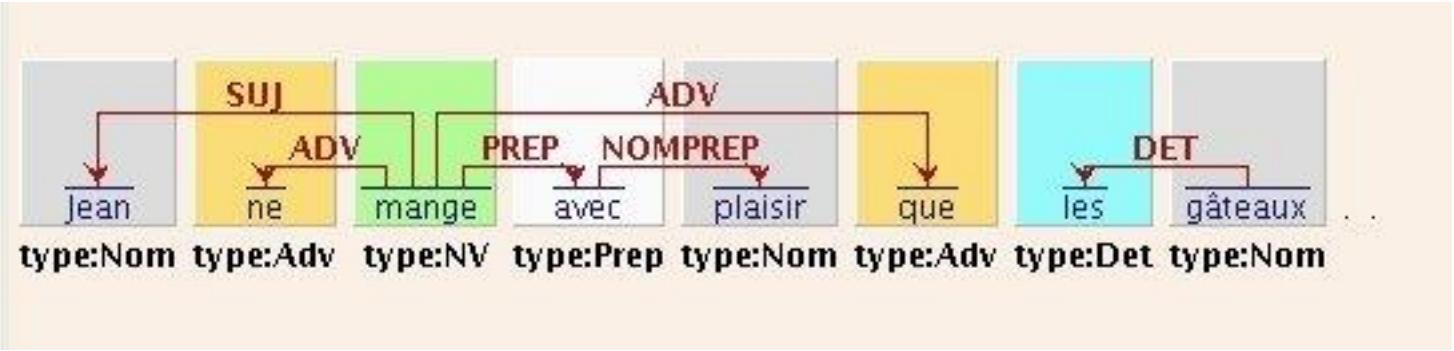
Analyse syntaxique

- Analyse syntaxique (au moins partielle) :
  - clauses : *When*-clause, *That*-clause (plusieurs types en anglais),
  - sub *Que*, sub conditionnel, sub relatif (pour le français, B. Habert)
  - Complexité des phrases (profondeur d'arbres syntaxiques)

- Difficultés : temps de calcul, taux d'erreur, ...

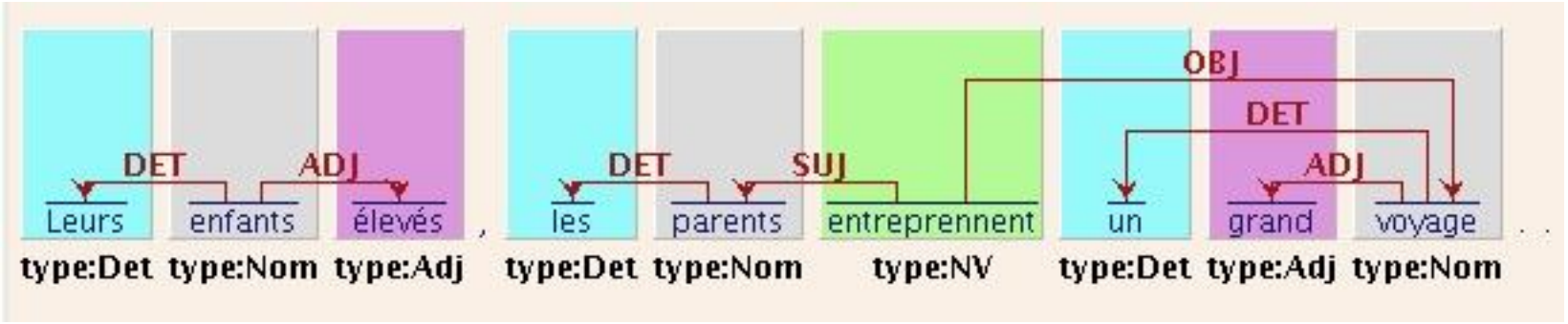
- Exemples de ressources :
  - ***SYNTEX*** (Didier Bourigault) : extraction de (certaines) dépendances syntaxiques
    - Analyse syntaxique robuste ;
    - Extraction et structuration d'un réseaux de syntagmes (nominaux, verbaux).

SYNTEX (Didier Bourigault) 1/3



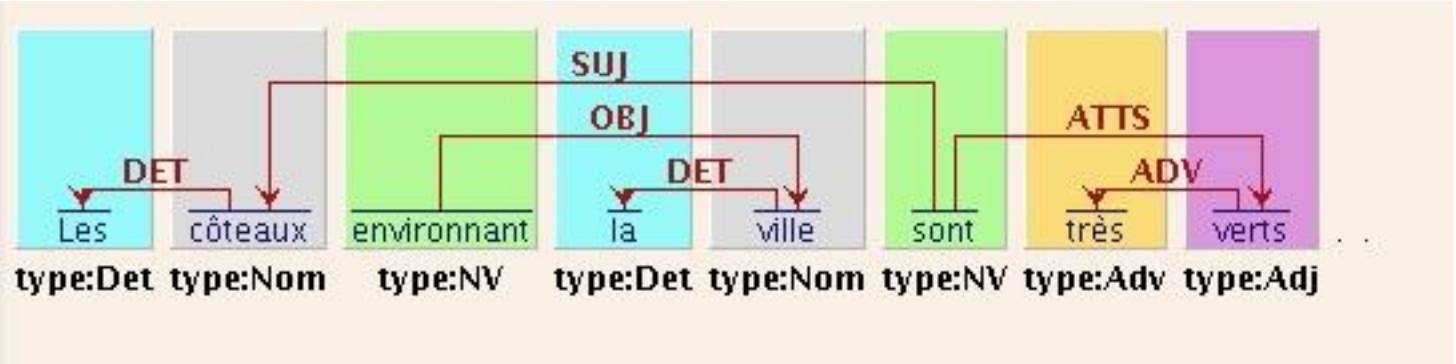
- analyse en dépendances
- dépendances entre tokens, pas de chunks préalables
- relations de dépendance autour du nom et autour du verbe
- un 'mot' → un seul recteur ; un recteur → plusieurs régis
- choix des relations de dépendance pragmatiques (références théoriques : Tesnière, Mel'čuk…)

SYNTEX (Didier Bourigault) 2/3



<SEQ id=2_2; analyse=1;>
<TXT>Leurs enfants élevés , les parents entreprennent un grand voyage .
<ETIQ>DetMP|leur|Leurs|1|DET;2| NomMP|enfant|enfants|2||DET;1,ADJ;3
AdjMP|élevé|élevés|3|ADJ;2| Typo|,|,|4|| DetMP|le|les|5|DET;6|
NomMP|parent|parents|6|SUJ;7|DET;5
VCONJP|entreprendre|entreprennent|7||SUJ;6,OBJ;10 DetMS|un|un|8|DET;10|
AdjMS|grand|grand|9|ADJ;10| NomMS|voyage|voyage|10|OBJ;7|DET;8,ADJ;9
Typo|.|.|11|| Typo|.|.|12||

SYNTEX (Didier Bourigault) 3/3



<SEQ id=2_4; analyse=1;>
<TXT>Les côteaux environnant la ville sont très verts .
<ETIQ>DetMP|le|Les|1|DET;2| NomInc|côteaux|côteaux|2|SUJ;6|DET;1
Ppr|environner|environnant|3||OBJ;5 DetFS|le|la|4|DET;5|
NomFS|ville|ville|5|OBJ;3|DET;4 VCONJP|être|sont|6||SUJ;2,ATTS;8
Adv|très|très|7|ADV;8| AdjMP|vert|verts|8|ATTS;6|ADV;7 Typo|.|.|9||
Typo|.|.|10||

Analyse sémantique

- classes sémantiques prédéfinies :
  - *verbes d'action , de communication, …*
- marqueurs spécifiques :
  - *Il est recommandé d'/de* + INF
  - *Pouvoir* Ind prés

- Exemples de ressources :
  - Listes de marqueurs spécifiques constituées à la main (ex.: marques de cohésion…)
  - *WordNet* (pour l'anglais), *Tropes* (commercialisé)…??
  - *Verbaction-Nomdaction* (libre)

Biber (2004) : traits mixtes issus de l'analyse morphosyntaxique, syntaxique, lexical et sémantique (quelques exemples…) 1/3

1. Pronouns and pro-verbs
2. Reduced forms and dispreferred structures
3. Prepositional phrases
4. Coordination
5. *WH*-Questions
6. Lexical specificity (*type/token* ratio, word length)
7. Nouns
nominalizations (ending in *-tion*, *-ment*, *-ness*, *-ity*)
nouns
7a. Semantic categories of nouns (animate, cognitive, quantity, place…Total : 8)
8. Verbs
8a. Tense and aspect markers
8b. Passives
8c. Modals

Douglas Biber (2004) "Conversation text types: A multi-dimensional analysis", *Proceedings of JADT'04.*

Biber (2004) : traits mixtes issus de l'analyse morphosyntaxique, syntaxique, lexical et sémantique (quelques exemples…) 2/3

8d. Semantic categories of verbs:

be as main verb
activity verb (*suggest, declare, tell*)
mental verb (*know, think, believe*)
causative verb (*let, assist, permit*)
occurrence verb *(increase, grow, become*)
existence verb (*possess, reveal, include*)
aspectual verb (*keep, begin, continue*)

Douglas Biber (2004) "Conversation text types: A multi-dimensional analysis", *Proceedings of JADT'04.*

Biber (2004) : traits mixtes issus de l'analyse morphosyntaxique, syntaxique, lexical et sémantique (quelques exemples…) 3/3

/…/

13. *That* complement clauses

13a. That clauses controlled by a verb

(Ex.: we predict that the water is here)

13b. That clauses controlled by a adjective

(Ex.: it is strange that he went there)

13c. That clauses controlled by a noun.

(Ex.: the proposal that he put forward was accepted)

Douglas Biber (2004) "Conversation text types: A multi-dimensional analysis", *Proceedings of JADT'04.*

Mise en place pour le texte en français : projet **TyPTex** (B. Habert *et al.*, 2000)

- Etiquetage morpho-syntaxique.
- Marquage typologique (regroupement, dégroupement, transformations, complémentations, omission…) :
  - Émergence de nouvelles catégories correspondant aux traits linguistiques dont on veut étudier la distribution…
- Outils/méthodes de comptages :
  - On construit la matrice des fréquences de chaque trait dans chaque texte ;
  - La matrice sert tant à la recherche des traits les plus pertinents pour une opposition qu'à la classification inductive ou supervisée.

Niveau typo-dispositionnel : format *html*

- HTML tags (28 chez M. Santini, …)
  - titres, listes, mails, images, tables…
  - liens (internes ou externes), types de liens suivant le suffixe (org, com..)
  - couleur, police de caractères…

Caractéristiques générales du document

- type/token ratio
- longueur du texte
- longueur des phrases
- …

M. Santini (2006) : jeux de traits mixtes 1/2

1_set

- The 50 most common words in English
- 24 POS tags
- 8 punctuation symbols (:;,!?' " ")
- 7 genre-specific facets (for the 7 web genre collection)
- 28 HTML tags
- 1 nominal attribute representing the length of the Web page (SHORT, MEDIUM and LONG)

2_set

- 100 POS trigrams
- 8 punctuation symbols
- genre-specific facets
- HTML tags
- 1 nominal attribute

Marina Santini (2006) "Some Issues in Automatic Genre Classification of Web Pages", in Proceedings of JADT'2006.

M. Santini (2006) : jeux de traits mixtes 1/2

- **3_set**
- 86 linguistic facets ;
- genre-specific facets ;
- 6 HTML facets ;
- 1 nominal attribute.

Exemple d'une matrice…

Problèmes de tels ensembles de traits

- Ils mettent à plat les traits récoltés à partir du texte

- Ne prennent pas en compte
  - *la concentration d'indices :*
    - Pour une même fréquence, besoin de distinguer des indices éparpillés dans le texte ou au contraire très fréquents dans un passage…

  - *la co-présence d'indices dans le texte :*
    - Verbes d'action (niveau sémantique) dans des listes (niveau typo-dispositionnel) ou dans des titres.

Bibliographie

- Une première synthèse en cours :

  http://www-lipn.univ-paris13.fr/~zimina/typologies/typo.html

- A lire, par exemple :

  Sven Meyer zu Eissen, Benno Stein Sven (2004). ”Genre Classification of Web Pages. User Study and Feasibility Analysis.” ***Proceedings of KI’2004: Advances in Artificial Intelligence.***