Étiquetage morpho-syntaxique

Benoît Habert

LIMSI – CNRS & université Paris X–Nanterre

http ://www.limsi.fr/Individu/habert - habert [@]
limsi.fr

Cours PluriTAL 12/01/2006

Plan

- **Exemples d'étiquetage morpho-syntaxique**
- Forme des étiquetages
- Ajuster traitement et données

Étiqueter : dimensions

[Ide & Romary 03]

- « Répartitoire » (*annotation scheme)* : ensemble des catégories distinguées ; définition en extension/intension
- Choix d'étiquettes (*encoding scheme*) : représentation de surface des catégories ('Nom commun masculin singulier' $\equiv$ NCMS)
- Mode d'annotation (*data architecture*) : réalisation « pratique »

Le Dormeur du Val

C’est un trou de verdure où chante une rivière
Accrochant follement aux herbes des haillons
D’argent ; où le soleil, de la montagne fière,
Luit : c’est un petit val qui mousse de rayons.

Un soldat jeune, bouche ouverte, tête nue,
Et la nuque baignant dans le frais cresson bleu,
Dort ; il est étendu dans l’herbe, sous la nue,
Pâle dans son lit vert où la lumière pleut.

Les pieds dans les glaïeuls, il dort. Souriant comme
Sourirait un enfant malade, il fait un somme :
Nature, berce-le chaudement : il a froid.

Les parfums ne font pas frissonner sa narine ;
Il dort dans le soleil, la main sur sa poitrine
Tranquille. Il a deux trous rouges au côté droit.

Etiquetage avec *Cordial* (1/2)

m o t	§	P h r a s e	d i a l o g u e	lemme	n b a m b i g.	Typegram	Code Gram	Codegram	S y n t a g m e
==== DEBUT DE PHRASE ====									
1	1	Le	1	le	A2	DETDMS	0xA000	D a - m s - d	2
1	1	Dormeur	2	dormeur	A2	NCMS	0xA000	N c m s	2
1	1	du	3	du		DETDMS	0xA000	D a - m s - d	4
1	1	Val	4	val		NCMS	0xA000	N c m s	4
==== FIN DE PHRASE ====									
==== DEBUT DE PHRASE ====									
\r\r									

Etiquetage avec *Cordial* (2/2)

P	...	N	P	Type	Sens
h	...	u	i	prop.	du
r	...	m	v		mot
a	...		o		
s	...	p	t		
e	...	r			
		o			
		p			
		.			
==== DEBUT DE PHRASE ====					
Le	...	1		Indépendante	
Dormeur	...	1		Indépendante	personne
du	...	1		Indépendante	
Val	...	1		Indépendante	
==== FIN DE PHRASE ====					
==== DEBUT DE PHRASE ====					
\r\r					

Etiquetage avec le *TreeTagger*

Le	DET :ART	le
Dormeur	NOM	dormeur
du	PRP :det	du
Val	NOM	val
C’	PRO :DEM	ce
est	VER :pres	être
un	DET :ART	un
trou	NOM	trou
de	PREP	de
verdure	NOM	verdure
où	PRO :REL	où
chante	VER :pres	chanter
une	DET :ART	un
rivière	NOM	rivière

Etiquetage avec *Unitex/Nooj/Intex*

{Pâle,pâle.A+z1 :ms :fs}
{dans,dans.PREP+z1}
{son,son.N+z1 :ms}
{son,son.DET+z1 :ms}
{lit,lire.V+z1 :P3s}
{lit,lit.N+z1 :ms}
{vert,vert.N+z1 :ms}
{vert,vert.A+z1 :ms}
{où,où.PRO+z1}
{la,le.PRO+z1 :3fs}
{la,le.DET+z1 :fs}
{la,la.N+z1 :ms :mp}
{lumière,lumière.N+z1 :fs}
{pleut,pleuvoir.V+z1 :P3s}

Morphologie avec *Flemm/Dérif* (1/2)

⋮

Dormeur NOM :Nc-s– dormeur [[dormir VER] eur NOM] (dormeur/NOM, dormir/VER) "(Agent habituel - Auteur exceptionnel - Instrument) de dormir"

⋮

chante VER(pres) :Vmip1s–1 chanter [chanter VER] (chanter/VER) "chanter"
chante VER(pres) :Vmip3s–1 chanter [chanter VER] (chanter/VER) "chanter"
chante VER(pres) :Vmmp2s–1 chanter [chanter VER] (chanter/VER) "chanter"
chante VER(pres) :Vmsp1s–1 chanter [chanter VER] (chanter/VER) "chanter"
chante VER(pres) :Vmsp3s–1 chanter [chanter VER] (chanter/VER) "chanter"

⋮

mousse NOM :Nc-s– mousse [mousse NOM] (mousse/NOM) "mousse"

⋮

tête NOM :Nc-s– tête [tête NOM] (tête/NOM) "tête"
nue NOM :Ncfs– nue [[nue ADJ] NOM] (nue/NOM, nue/ADJ) "Entité dont la propriété vue comme saillante est d'être nue"

⋮

Morphologie avec *Flemm/Dérif* (2/2)

⋮

⋮

Dort VER(pres) :Vmip3s–3 dormir [dormir VER] (dormir/VER) "dormir"

⋮

étendu VER(pper) :Vmps-sm– étendre [é [tendre VER] VER] (étendre/VER, tendre/VER) "tendre jusqu'au bout"

⋮

nue NOM :Ncfs– nue [[nue ADJ] NOM] (nue/NOM, nue/ADJ) "Entité dont la propriété vue comme saillante est d'être nue"

⋮

Etiquetages...

Cordial (30 étiquettes réalisées sur 309)

ADJFS (3 o.), ADJINV (3 o.), ADJMIN (1 o.), ADJMS (4 o.), ADJNUM (1 o.), ADJPIG (1 o.), ADJSIG (4 o.), ADV (4 o.), COO (1 o.), DETDFS (6 o.), DETDMS (6 o.), DETDPIG (5 o.), DETIFS (1 o.), DETIMS (5 o.), DETPOSS (3 o.), NCFP (1 o.), NCFS (13 o.), NCMP (5 o.), NCMS (10 o.), NCSIG (1 o.), NHMIN (1 o.), PCTFAIB (12 o.), PCTFORTE (12 o.), PDS (2 o.), PREP (11 o.), PRI (4 o.), SUB (1 o.), VINF (1 o.), VPARPMS (1 o.), VPARPRES (3 o.)

TreeTagger (19 étiquettes réalisées sur 59)

ADJ (11 o.), ADV (5 o.), DET :ART (19 o.), DET :POS (3 o.), KON (1 o.), NOM (33 o.), PRO :DEM (2 o.), PRO :PER (6 o.), PRO :REL (4 o.), PRP (11 o.), PRP :det (3 o.), PUN (18 o.), SENT (5 o.), VER :aux :pres (1 o.), VER :cond (1 o.), VER :infi (1 o.), VER :pper (1 o.), VER :ppre (3 o.), VER :pres (12 o.)

Comparaison d'étiquetages (1/2)

v.	*Cordial*				*TreeTagger*				
	mot	*lemme*	*cat.*	*POS*	*mot*	*lemme*	*cat.*		*POS*
		⋮				⋮			
1	chante	chanter	VIND-P3S	V	chante	chanter	VER pres	:	V
1	une	un	DETIFS	D	une	un	DET ART	:	D
1	rivière	rivier	NCFS	N	rivière	**rivière**	NOM		N
		⋮				⋮			
3	argent	argent	ADJINV	A	argent	argent	**NOM**		N
		⋮				⋮			
4	mousse	**mousser**	**VIND-P3S**	V	mousse	mousse	NOM		N
		⋮				⋮			
5	tête	tête	NCFS	N	tête	tête	NOM		N
5	nue	**nu**	**ADJFS**	A	nue	nue	NOM		N
		⋮				⋮			

Comparaison d'étiquetages (2/2)

v.	*Cordial*				*TreeTagger*				
	mot	*lemme*	*cat.*	*POS*	*mot*	*lemme*	*cat.*		*POS*
		⋮				⋮			
9	comme	**comme**	**SUB**	C	comme	comme	ADV		R
		⋮				⋮			
11	Nature	nature	ADJINV	A	Nature	nature	**NOM**		N
11	berce	bercer	**VIMP-P2S**	V	berce	bercer	VER pres	:	V
		⋮				⋮			

Plan

- Exemples d'étiquetage morpho-syntaxique
- **Forme des étiquetages**
- Ajuster traitement et données

Conventions implicites *vs.* explicites

TreeTagger

Le	$\longrightarrow$	DET :ART	$\longrightarrow$	le¶
Dormeur	$\longrightarrow$	NOM	$\longrightarrow$	dormeur¶
du	$\longrightarrow$	PRP :det	$\longrightarrow$	du¶
Val	$\longrightarrow$	NOM	$\longrightarrow$	val¶

Les sorties de TreeTagger obéissent à un *format délimité*, du type des tables de SGBD (csv : *comma separated values*) : une entité $\equiv$ une ligne (**n** ou **r***n)** ; une facette $\equiv$ une colonne (**t**)

	forme	*catégorie*	*lemme*
mot_1	Le	DET : ART	le
mot_2	Dormeur	NOM	dormeur
mot_3	du	PRP :det	du
mot_4	Val	NOM	val

Notation idiosyncrasique

m	§	P	d	lemme	n	Typegram	Code	Codegram	S
o		h	i		b		Gram		y
t		r	a						n
		a	l		a				t
		s	o		m				a
		e	g		b				g
			u		i				m
			e		g.				e
==== DEBUT DE PHRASE ====									
1	1	Le	1	le	A2	DETDMS	0xA000	D a - m s - d	2
1	1	Dormeur	2	dormeur	A2	NCMS	0xA000	N c m s	2
1	1	du	3	du		DETDMS	0xA000	D a - m s - d	4
1	1	Val	4	val		NCMS	0xA000	N c m s	4
==== FIN DE PHRASE ====									
==== DEBUT DE PHRASE ====									
\r\r									

Transformation *ad hoc*

m	*§*	***P***	*d*	***lemme***	*n*	***Typegram***	*Code*	***Codegram***	*S*
o		***h***	*i*		*b*		*Gram*		*y*
t		***r***	*a*						*n*
		a	*l*		*a*				*t*
		s	*o*		*m*				*a*
		e	*g*		*b*				*g*
			u		*i*				*m*
			e		*g.*				*e*
==== DEBUT DE PHRASE ====									
1	*1*	**Le**	*1*	**le**	*A2*	**DETDMS**	*0xA000*	**D a - m s - d**	*2*
1	*1*	**Dormeur**	*2*	**dormeur**	*A2*	**NCMS**	*0xA000*	**N c m s**	*2*
1	*1*	**du**	*3*	**du**		**DETDMS**	*0xA000*	**D a - m s - d**	*4*
1	*1*	**Val**	*4*	**val**		**NCMS**	*0xA000*	**N c m s**	*4*
==== FIN DE PHRASE ====									
==== DEBUT DE PHRASE ====									
\r\r									

Format *vs* catégories

- TreeTagger

<w lemma="le" type="DET :ART">Le</w>
<w lemma="dormeur" type="NOM">dormeur</w>
<w lemma="du" type="PRP :det">du</w>
<w lemma="val" type="NOM">val</w>

- Cordial (jeu$_1$)

<w lemma="le" type="DETDMS">Le</w>
<w lemma="dormeur" type="NCMS">dormeur</w>
<w lemma="du" type="DETDMS">du</w>
<w lemma="val" type="NCMS">val</w>

- Cordial (jeu$_2$)

<w lemma="le" type="Da-ms-d">Le</w>
<w lemma="dormeur" type="Ncms">dormeur</w>
<w lemma="du" type="Da-ms-d">du</w>
<w lemma="val" type="Ncms">val</w>

Choix dans les notations

- Notation idiosyncrasique *vs.* homologuée (*de facto* ou norme)
- Notation implicite *vs.* explicite
- Notation validable *vs.* sans validation explicite
- Notation traitant un niveau / plusieurs niveaux d'annotation
- Notation à plat | enchâssante | enchevêtrée

Plan

- Exemples d'étiquetage morpho-syntaxique
- Forme des étiquetages
- **Ajuster traitement et données**

Ajuster traitement et données (1/2)

- Régler un instrument / adapter les données
  - Exemple : *Dormeur* en phrases (*vs* vers)

Le dormeur du val
C’ est un trou de verdure où chante une rivière accrochant follement aux herbes des haillons d’ argent ; où le soleil , de la montagne fière , luit : c’ est un petit val qui mousse de rayons . · · · Les parfums ne font pas frissonner sa narine ; il dort dans le soleil , la main sur sa poitrine tranquille . Il a deux trous rouges au côté droit .

Ajuster traitement et données (2/2)

vers

...

rivière	rivier	NCFS
===== FIN DE PHRASE =====		
==== DEBUT DE PHRASE ====		
Accrochant	accrocher	VPARPRES
follement	follement	ADV
aux	au	DETDPIG
herbes	herbe	NCFP
des	de	DETDPIG
haillons	haillon	NHMIN
===== FIN DE PHRASE =====		
==== DEBUT DE PHRASE ====		
D’	de	PREP
argent	argent	**ADJINV**
;	;	PCTFORTE
===== FIN DE PHRASE =====		

phrase

...

rivière	rivier	NCFS
accrochant	accrochant	VPARPRES
follement	follement	ADV
aux	au	DETDPIG
herbes	herbe	NCFP
des	de	DETDPIG
haillons	haillon	NHMIN
d’	de	PREP
argent	argent	**NCMS**
;	;	PCTFORTE
===== FIN DE PHRASE =====		