Documentation Syntex
v 1.2
D. Bourigault
didier.bourigault@univ-tlse2.fr

http://w3.univ-tlse2.fr/erss/membres/bourigault/

Septembre 2007

1. Conditions requises

Il faut :

- perl v5.8.4 (minimum) built for i386-linux-thread-multi
- le module perl LWP, fonction MD5 (à chaque lancement, un message de "Vérification licence" est envoyé à un serveur)
- le Treetagger installé , avec en particulier dans le répertoire d'installation du Treetagger :
  - lib/french.par : fichier de paramètres
  - /bin/tree-tagger : module exécutable

  http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html

2. Format du fichier d'entrée (.txt)

Le fichier d'entrée est un fichier texte avec l'extension .txt (ex. : monfich.txt).
Il ne comporte pas de retour-chariots au sein des phrases (le module de découpage en phrases considère le retour-chariot comme un délimiteur de phrase).

Si le corpus à analyser est constitué de plusieurs parties, il est possible de les marquer avec des balises de la forme <#iddoc> :

- la balise est sur une ligne (en début de ligne)
- elle ne contient pas de _ et de ;

Le module de découpage en séquences (phrases) affecte à chaque séquence un identifiant qui a la forme *iddoc_n*, où *n* est le numéro de la phrase dans la partie dont l'identifiant est *iddoc* (cf. section 5).

3. La chaîne d'analyse syntaxique

Le système Syntex est une chaîne d'analyse syntaxique organisée en 4 phases :

1- *Préétiquetage* : ce module effectue la segmentation en phrases, la tokénisation en mots et le préétiquetage des mots ; ce module et les ressources qu'il utilise ont été développés par la société Synomia.

2- *Etiquetage morphosyntaxique* : l'étiquetage est effectué par l'outil Treetagger de l'Université de Stuttgart.

3- *Conversion* : ce module effectue la conversion des sorties du Treetagger aux entrées attendues par l'analyseur Syntex ; il a été développé conjointement par l'ERSS et la société Synomia.

4- *Analyse syntaxique* : l'analyse syntaxique est réalisée par l'analyseur Syntex ; il a été développé par l'ERSS.

4. Lancement

Dans lanceSyntex.sh, modifer les variables d'environnement DIRBASE et DIRTT.

Dans le répertoire courant, le fichier d'entrée a l'extension txt (ex. : monfich.txt). Lancer :

```
$DIRBASE/lanceSyntex.sh monfich
```

Les fichiers intermédiaires suivants sont créés :

- test.tok : résultat de la phase 1 de préétiquetage
- test.tag : résultat de la phase 2 d'étiquetage
- test.2syntex : résultat de la phase 3 de conversion
- test.anasynt : résultat de la phase 4 d'analyse syntaxique (fichier au format « syntex »)
- test-syntex.xml : résultat de la phase 4 d'analyse syntaxique (fichier au format xml)

5. Format du fichier de sortie .anasynt

Ce fichier contient les résultats de l'analyse syntaxique de chacune des séquences du corpus. Pour chaque séquence, 3 lignes : SEQ, TXT, ETIQ

<SEQ id=iddoc_n; analyse=1>
L'identifiant de séquence est la concaténation de l'identifiant de document figurant dans le fichier initial (balise <#iddoc>) et du numéro de la séquence dans le document (n) (cf. section 3).

<TXT>…
La séquence. Celle-ci est donnée telle qu'elle est reconstituée après tokénisation et étiquetage.

<ETIQ>cat|lemme|forme|num|gouverneur|dependant
Sur cette ligne figure le résultat de l'analyse syntaxique de la phrase. Pour chacun des mots (tokens), séparés par des tabulations, on trouve les informations suivantes (le séparateur est |) :

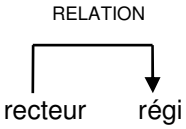
- **cat** : sa catégorie morphosyntaxique. La liste des catégories est donnée dans le tableau 1.
- **lemme** : son lemme (masculin pour les noms, masculin singulier pour les adjectifs, infinitif pour les verbes, ...)
- **forme** : la forme sous laquelle apparaît le mot dans la phrase
- **num** : le rang du mot dans la séquence
- **gouverneur** (RELATION;numgouverneur) : le couple (relation de dépendance syntaxique ; numéro du gouverneur). Le séparateur entre la relation et le numéro du gouverneur est le point-virgule. La liste des relations de dépendance syntaxique est donnée dans le tableau 2. Un mot ne peut avoir qu'un seul gouverneur. Ce champ sert aussi à indiquer les relations d'antécédence relative (REL) entre un mot et le pronom relatif dont il est l'antécédent. Dans le cas des pronoms relatifs, il peut donc y avoir 2 couples dans ce champ (séparés alors par une virgule).
- **dependant** ($RELATION_1$;$numdependant_1$, …) : la liste des numéros des dépendants avec la relation grammaticale. Un mot peut avoir plusieurs dépendants. Le séparateur entre la relation et le numéro d'un dépendant est le point-virgule. Le séparateur entre les différents couples est la virgule.

L'information sur la relation de dépendance entre un dépendant et un gouverneur est donc donnée 2 fois : dans le champ dépendant du gouverneur, et dans le champ gouverneur du dépendant.

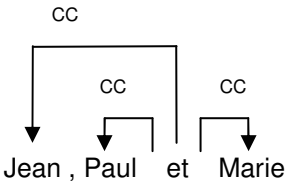
Quelques jolis dessins valent mieux que de longs discours[1], illustrations :

[1] Pour une description détaillée de l'analyseur Syntex, voir mon mémoire d'Habilitation à Diriger les Recherches (http://w3.univ-tlse2.fr/erss/membres/bourigault/hdr.html)

Modèle général

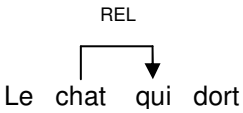


Coordination : CC

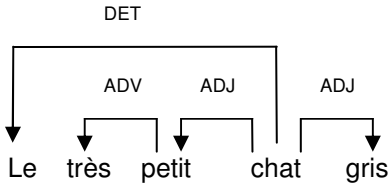



```
NomPrXXPrenom|Jean|Jean|1|CC;4|  TypoNomPr|,|,|2||  NomPrXXPrenom|Paul|Paul|3|CC;4|
CCoordNomPr|et|et|4||CC;1,CC;3,CC;5  NomPrXXPrenom|Marie|Marie|5|CC;4|
```

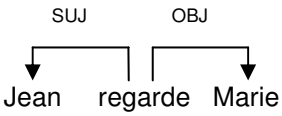
Antécédence relative : REL



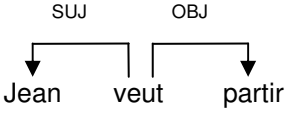

```
Det??|le|le|5|DET;6|  NomMS|chat|chat|6|REL;7|DET;5 ProRel|qui|qui|7|SUJ;8|REL;6
VCONJS|dormir|dort|8||SUJ;7
```



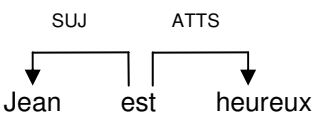

```
Det??|le|Le|1|DET;4|  Adv|très|très|2|ADV;3|Adj?S|petit|petit|3|ADJ;4|ADV;2
NomMS|chat|chat|4||DET;1,ADJ;3,ADJ;5 Adj??|gris|gris|5|ADJ;4|
```



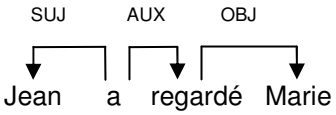

```
NomPrXXPrenom|Jean|Jean|1|SUJ;2|     VCONJS|regarder|regarde|2||SUJ;1,OBJ;3
NomPrXXPrenom|Marie|Marie|3|OBJ;2|
```



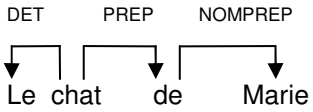

```
NomPrXXPrenom|Jean|Jean|1|SUJ;2|     VCONJS|vouloir|veut|2||SUJ;1,OBJ;3
VINF|partir|partir|3|OBJ;2|
```



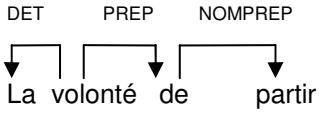

```
NomPrXXPrenom|Jean|Jean|1|SUJ;2| VCONJS|être|est|2||SUJ;1,ATTS;3
AdjM?|heureux|heureux|3|ATTS;2|
```



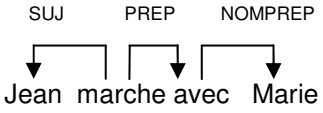

```
NomPrXXPrenom|Jean|Jean|1|SUJ;2|     VCONJS|a|avoir|2||SUJ;1,AUX;3
      PpaMS|regardé|regarder|3|AUX;2|OBJ;4 NomPrXXPrenom|Marie|Marie|4|OBJ;3|
```



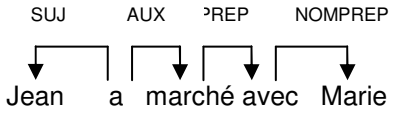

```
Det??|le|Le|1|DET;2|  NomMS|chat|chat|2||DET;1,PREP;3
Prep|de|de|3|PREP;2|NOMPREP;4  NomPrXXPrenom|Marie|Marie|4|NOMPREP;3|
```



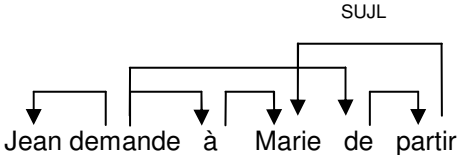

```
Det??|la|La|1|DET;2|  NomFS|volonté| volonté |2||DET;1,PREP;3
Prep|de|de|3|PREP;2|NOMPREP;4  VINF|partir|partir|4|NOMPREP;3|
```




```
NomPrXXPrenom|Jean|Jean|1|SUJ;2|  VCONJS|marcher|marche|2||SUJ;1,PREP;3
Prep|avec|avec|3|PREP;2|NOMPREP;4  NomPrXXPrenom|Marie|Marie|4|NOMPREP;3|
```




```
NomPrXXPrenom|Jean|Jean|1|SUJ;2|  VCONJS|a|avoir|2||SUJ;1,AUX;3
PpaMS|marché|marcher|3|AUX;2|PREP;4  Prep|avec|avec|4|PREP;3|NOMPREP;5
NomPrXXPrenom|Marie|Marie|5|NOMPREP;4|
```




```
NomPrXXPrenom|Jean|Jean|1|SUJ;2|  VCONJS|demande|demander|2||SUJ;1,PREP;3,PREP;5
Prep|à|à|3|PREP;2|NOMPREP;4  NomPr|Marie|Marie|4|NOMPREP;3,SUJL;6|
Prep|de|de|5|PREP;2|NOMPREP;6|  VINF|partir|partir|6|NOMPREP;5|SUJL;4
```

Tableau 1 : liste des catégories morphosyntaxiques

	Catégorie	Définition	nb (corpus 1M mots)
Adj			
	Adj??	adjectif de genre et nombre indéterminés	180
	Adj?P	adjectif de genre indéterminé et de nombre pluriel	5914
	Adj?S	adjectif de genre indéterminé et de nombre singulier	1179
	AdjFP	adjectif de genre féminin et de nombre pluriel	7410
	AdjFS	adjectif de genre féminin et de nombre singulier	18082
	AdjMP	adjectif de genre masculin et de nombre pluriel	9928
	AdjMS	adjectif de genre masculin et de nombre singulier	22347
Adv			
	Adv	adverbe	39827
	AdvGP	groupe prépositionnel adverbial	7379
CCoord			
	CCoord	Conjonction de coordination	9511
	CCoordCAT	Conjonction de coordination (où CAT est la catégorie des éléments coordonnés, quand la relation de coordination a été identifiée)	12335
CSub			
	CSub	conjonction de coordination	10782
Det			
	Det	déterminant	7863
	Det??	déterminant de genre et nombre indéterminés	2059
	DetFP	déterminant de genre féminin et de nombre pluriel	7936
	DetFS	déterminant de genre féminin et de nombre singulier	39095
	DetMP	déterminant de genre masculin et de nombre pluriel	13518
	DetMS	déterminant de genre masculin et de nombre singulier	53772
	DetNum	déterminant numérique	4816
Elim			
	Elim	catégorie éliminatoire	436
Nom			
	Nom??	nom de genre et nombre indéterminés	31
	Nom?P	nom de genre indéterminé et de nombre pluriel	36
	Nom?S	nom de genre indéterminé et de nombre singulier	37
	NomFP	nom de genre féminin et de nombre pluriel	20478
	NomFS	nom de genre féminin et de nombre singulier	61198
	NomInc	nom inconnu	4019
	NomMP	déterminant de genre masculin et de nombre pluriel	30866
	NomMS	déterminant de genre masculin et de nombre singulier	69915
NomPr			
	NomPr	nom propre	18941
	NomPrXXInc	nom propre inconnu (n'appartenant pas aux dictionnaires)	27028
	NomPrXXPrenom	prénom	9938
Nom			

	NomXXAdr	adresse	68
	NomXXDate	date	12543
	NomXXHeure	heure	288
	NomXXMes	mesure	1554
	NomXXMon	monnaie	1157
	NomXXNum	numérique	5587
	NomXXUrl	url	4
Ppa			
	PpaFP	participe passé de genre féminin et de nombre pluriel	2667
	PpaFS	participe passé de genre féminin et de nombre singulier	5343
	PpaMP	participe passé de genre masculin et de nombre pluriel	5244
	PpaMS	participe passé de genre masculin et de nombre singulier	19743
	PpaMSp	participe passé passif (été)	1718
Ppr			
	Ppr	participe présent	3706
	Pprp	participe présent passif	50
Prep			
	Prep	préposition	143635
PrepDet			
	PrepDet	"de" non désambiguïsé Det ou Prep)	1999
Pro			
	Pro	pronom personel	39821
ProRel			
	ProRel	pronom relatif	11509
Typo			
	Typo	marque de typographie	140473
	TypoCoordCAT		1956
VCONJ			
	VCONJP	verbe conjugué au pluriel	15011
	VCONJPp	verbe conjugué au pluriel au passif	1113
	VCONJS	verbe conjugué au singulier	44358
	VCONJSp	verbe conjugué au singulier au passif	2045
VINF			
	VINF	verbe à l'infinitif	20791
	VINFp	verbe à l'infinitif au passif	762

Tableau 2 : liste des relations de dépendance

Un certain nombre de cas ne sont pas encore traités de façon satisfaisante. Il sont marqués :-(

Relation	cat gouv.	cat dép.	exemple	nb (corpus 1M mots)
ADJ				
	Nom	Adj	chat gris	50973
		Ppa	chat allongé	7420
		Ppr	chat miaulant	1170
	NomPr	Adj	Jean, heureux d'être là :-(	1057
		Ppa	Jean, allongé sur le canapé :-(	75
		Ppr	Jean, observant le chat :-(	19
	Pro	Adj	celui, heureux :-(	12
		Ppa	celui vu :-(	40
		Ppr	celui allant :-(	18
ADV				
	Adj	Adv	très rapide	4863
	Adv		très rapidement	1784
	Det		environ 10000	104
	Nom		ex- président	283
	NomPr		ex- KGB	135
	Ppa		souvent vu	4424
	Ppr		voyant souvent	566
	VCONJ		il court vite	18718
		Nom	il dort le matin	9
	VINF	Adv	il peut courir vite	1915
		Nom	il peut dormir le matin	0
APPOS				
	Nom	Adj	le chat, joyeux	955
		Ppa	le chat, allongé	1684
		Ppr	le chat, dormant	637
	NomPr	Adj	Mistigri, joyeux	75
		Ppa	Mistigri, allongé	124
		Ppr	Mistigri, dormant	38
	Pro	Adj	celui, joyeux	11
		Ppa	celui, allongé	11
		Ppr	celui, dormant	1
ATTO				
	Ppa	Adj	il l'a rendu joyeux	71
		Nom	il l'a nommé directeur	65
		NomPr	il l'a surnommé Milou	4
		Ppa	il l'a rendu énervé	21
	Ppr	Adj	rendant joyeux le chat	12
		Ppa	rendant énervé le chat	3
	VCONJ	Adj	il le rend joyeux	86
		Nom	il le nomme directeur	3
		NomPr	il le surnomme Milou	0
		Ppa	il le rend énervé	52
	VINF	Adj	rendre joyeux le chat	30
		Nom	être nommé directeur	0
		NomPr	le surnommer Milou	0
		Ppa	rendre énervé le chat	5
ATTS				
	Ppa	Adj	a été gris	185
		Nom	a été le chat	311
		NomPr	a été Mistigri	5
		Ppa	a semblé énervé	17
		Pro	il est devenu celui	11
		ProRel	qu'ils ont été	12

		VINF	a semblé dormir	10
	Ppr	Adj	étant gris	22
		Nom	étant le chat	20
		NomPr	étant Mistigri	1
		Ppa	semblant énervé	0
		Pro	devenant celui	6
		VINF	semblant dormir	1
	VCONJ	Adj	est gris	2708
		Nom	est le chat	3300
		NomPr	est Paris	127
		Ppa	semble énervé	146
		Ppr	:-(est intéressant	18
		Pro	est celui	258
		ProRel	que sont les chat	202
		VINF	semble dormir	274
	VINF	Adj	être gris	236
		Nom	étant le chat	313
		NomPr	être Paris	11
		Ppa	sembler énervé	7
		Ppr	:-(être intéressant	8
		Pro	peut être celui	11
		ProRel	que peuvent être les chat	11
		VINF	sembler dormir	3
AUX				
	Ppa	Ppa	a été vu	1688
	Ppr		étant vu	184
	VCONJ		a vu	17176
	VINF		avoir vu	1312
COMP				
	CSub	Adj	aussi malin que rapide	23
		Adv	plus que souvent	28
		Nom	autre que le chat	153
		NomPr	moins que Paris	16
		Ppa	aussi rapide qu'énervé	0
		Prep	moins qu'avec le chat	31
		Pro	moins que celui	54
		VCONJ	vouloir qu'il soit	8624
		VINF	que boire	0
CPL				
	Adj	CSub	aussi grand que	97
	Adv		plus que	366
	Nom		le fait que	140
	NomPr		:-(	0
	Ppa		plus énervé que	17
	Ppr		:-(	0
	Prep		autant à Marie qu'à Jean	1
	PrepDet		:-(	0
	Pro		:-(tels que	15
	VCONJ		c' est ici que	36
	VINF		:-(	7
DET				
	Nom	Det	le chat	120970
		Prep	n'a pas vu de chat	291
	NomPr	Det	la France	6642
	Pro		le mien, la nôtre	29
EPI				
	Nom	Nom	le coin cuisine	5492
		NomPr	le chat Mistigri	4719
		Pro	:-(	1
	NomPr	Nom	Toulouse II	664
		NomPr	San Antonio	3848

I_ADJ				
	Typo	Typo	, rapide ,	374
I_ADV				
	Typo	Typo	, lentement ,	1721
I_PREP				
	Typo	Typo	, avec le chat ,	0
NNPR				
	NomPr	Nom	Monsieur Dupont	0
		NomPr	Jean Dupont	10108
NOMPREP				
	Prep	Adv	:-(d'autant	10
		Nom	avec le chat	104346
		NomPr	avec Marie	18297
		Ppr	en mangeant	1029
		Pro	avec lui	1652
		ProRel	pour lequel	597
		VINF	pour prendre	12359
	PrepDet	Nom	du chat	1832
		NomPr	du Liban	109
		VINF	de la voir	1
OBJ				
	Ppa	CSub	a vu que	675
		Nom	a vu le chat	4747
		NomPr	a vu Marie	200
		Pro	il a vu quelqu'un	57
		ProRel	qu'il a vu	258
		VINF	a voulu manger	814
	Ppr	CSub	voyant que	178
		Nom	voyant le chat	1571
		NomPr	voyant Marie	74
		Pro	le voyant	81
		VINF	voulant manger	100
	VCONJ	CSub	il voit que	1687
		Nom	il voit le chat	10953
		NomPr	il voit Marie	579
		Pro	il le voit	1590
		ProRel	le chat qu'il voit	807
		VINF	il veut voir	5198
	VINF	CSub	voir que	504
		Nom	voir le chat	9280
		NomPr	voir Marie	339
		Pro	le voir	834
		ProRel	que je tente de voir	206
		VINF	laisser faire	606
OBJ1				
	Ppa	Nom	j'ai vu le chat manger	39
		NomPr	j'ai vu Marie manger	4
		ProRel	que j'ai vu manger	5
	Ppr	Nom	en voyant le chat manger	9
		NomPr	en voyant Marie manger	0
		Pro	en la voyant manger	1
	VCONJ	Nom	je vois le chat manger	72
		NomPr	je vois Marie manger	5
		Pro	je la vois manger	40
		ProRel	que je vois manger	21
	VINF	Nom	voir le chat manger	73
		NomPr	voir Maire manger	11
		Pro	la voir manger	26
PAR				
	Typo	Typo	( )	5987
PREP				

	Adj	Prep	facile à	2612
	Nom		chat de	66157
		ProRel	le chat duquel	36
	NomPr	Prep	Afrique du Sud	703
	Ppa		équipé de	18287
		ProRel	à qui il a donné	302
	Ppr	Prep	regardant vers	1691
		Pro	lui donnant	23
	Pro	Prep	celui de	769
	VCONJ		il mange avec	16029
		Pro	je lui donne	421
		ProRel	auquel je tiens	654
	VINF	Prep	manger	8232
		Pro	pour lui donner	103
		ProRel	auquel je veux parler	48
REF				
	Ppr	Pro	se voyant	227
	VCONJ		il se voit	5144
	VINF		se voir	1770
REL				
	Nom	ProRel	l'homme qui	7221
	NomPr		Marie qui	965
	Pro		celui qui	1292
SUJ				
	VCONJ	Nom	le chat mange	24002
		NomPr	Marie Mange	5663
		Pro	il mange	19269
		ProRel	qui mange	6600
		VINF	dormir est	125
SUJL				
	VINF	Nom	permettre au chat de manger	1262
		NomPr	permettre à Marie de manger	389
		Pro	lui permettre de manger	956
		ProRel	auquel on a permis de manger	304