



## CONVERSIÓ FORMAT WIKI A XML

Col·leccions WikiXML  
(<http://lps.science.uva.nl/WikiXML/>)

Nombre d'entrades: 159.716.  
Nombre de tokens: 49.876.432.  
Versió de la Viquipèdia: 23/01/2009  
URL: <http://www.glicom.upf.edu/publicacions/recursos>  
Llicència: GNU Free Documentation License  
Disponibilitat: abril, 2009.

## TOKENITZACIÓ I LEMATITZACIÓ

TreeTagger

## RECUPERACIÓ D'ENLLAÇOS

```
diverses JQ divers 0
vil·les N5 vil·la 0
romanes JQ romà 0
pels N5 pels0
volts N5 volt 0
de P de 0
Terrassa N5 terrassa 0
, Fc , 0
concretament D4 concretament 0
a P a 0
Ca N5 ca 0
n' VJ n' 0
Amat VC amar 0
, Fc , 0
l' N4 l' 0
Aiguacuit N5 aiguacuit 0
, Fc , 0
Ca N5 ca B-/wiki/Ca_n%27Anglada
n' VJ n' I-/wiki/Ca_n%27Anglada
Anglada N4 anglada I-/wiki/Ca_n%27Anglada
, Fc , 0
Can N5 can 0
Fatjó N4 Fatjó 0
```

## PART OF SPEECH, MULTIWORDS

FreeLing

```
diverses divers DIOFP0
vil·les vil·les NCFP000
romanes romà 02696998 AQ0FP0
per per SPS00
els el DA0MP0
volts volt NCMP000
de de SPS00
Terrassa terrassa 06726901 NP00000
, , Fc
concretament concretament RG
a a SPS00
Ca ca NP00000
n' en PP3CN000
Amat amat NP00000
, , Fc
l' el DA0CS0
Aiguacuit aiguacuit NP00000
, , Fc
Ca ca NP00000
n' en PP3CN000
Anglada anglada NP00000
, , Fc
```

## PART OF SPEECH

TreeTagger

## SuperSense Tagger (Average perceptron)

(Ciaramita and Altun, 2006)

## JNET (CRF)

([www.julielab.de](http://www.julielab.de))

## DeSR

(parser de dependències)

(Attardi et al., 2007)

```
diverses O
vil·les O
romanes O
pels O
volts O
de O
Terrassa B-LOC
, O
concretament O
a O
Ca B-LOC
n' I-LOC
Amat I-LOC
, O
l' O
Aiguacuit B-MISC
, O
Ca B-LOC
n' I-LOC
Anglada I-LOC
, O
Can B-LOC
Fatjó I-LOC
, O
Can B-LOC
Poal I-LOC
, O
les O
Martines B-LOC
, O
Sant B-LOC
Pere I-LOC
```

```
diverses JQ divers O -1.0
vil·les N5 vil·la O -1.0
romanes JQ romà O -1.0
pels N5 pels O -1.0
volts N5 volt O -1.0
de P de O -1.0
Terrassa N5 terrassa B-LOC 0.9881769385696958
, Fc , O -1.0
concretament D4 concretament O -1.0
a P a O -1.0
Ca N5 ca B-LOC 0.943460222154275
n' VJ n' I-LOC 0.9179944688854016
Amat VC amar I-LOC 0.9179944688854016
, Fc , O -1.0
l' N4 l' O -1.0
Aiguacuit N5 aiguacuit B-LOC 0.5136405835076326
, Fc , O -1.0
Ca N5 ca B-LOC 0.8792401887993582
n' VJ n' I-LOC 0.8645079451672626
Anglada N4 anglada I-LOC 0.8645079451672626
, Fc , O -1.0
Can N5 can B-LOC 0.9208265906286328
Fatjó N4 Fatjó I-LOC 0.9203308820570639
, Fc , O -1.0
Can N5 can B-LOC 0.960148528302626
```

```
diverses di divers 23 ESPEC
vil·les nc vil·les 21 SN
romanes aq romà 23 SADJ
per sp per 13 CC
els da el 27 ESPEC
volts nc volt 25 SN
de sp de 27 SP
Terrassa np terrassa 28 SN
, Fc , 25 PUNC
concretament rg concretament 32 ADJUNCT
a sp a 25 CONJUNCT
Ca np ca 32 SN
n' pp en 33 SN
Amat np amat 34 SN
, Fc , 38 PUNC
l' da el 38 ESPEC
Aiguacuit np aiguacuit 35 SN
, Fc , 38 PUNC
```

**SuperSenseTagger**  
Nombre total de NE: 4.527.594  
Lloc: 1.295.100  
Organització: 857.546  
Persona: 1.472.553  
Miscel·lània: 902.395

**JNET**  
Nombre total de NE: 3.527.785  
Lloc: 1.186.978  
Organització: 641.036  
Persona: 1.297.330  
Miscel·lània: 402.441

## TRENCAMENT DE MW I UNIÓ DE PREPOSICIONS CONTRACTES

## ALINEACIÓ

FORMA	SST NE	JNET PosTag	LEMA	NE C.SCORE	DeSR+FreeLing PosTag	LEMA	Depen.	Funció	WNSense	PosTag	Rec. Enllaços Enllaç
volts	O	N5	volt	O -1.0	nc	volt	25	SN		NCMP000	0
de	O	P	de	O -1.0	sp	de	27	SP		SPS00	0
Terrassa	B-LOC	N5	terrassa	B-LOC 0.98	np	terrassa	28	SN	06726901	NP00000	0
,	O	Fc	,	O -1.0	Fc	,	25	PUNC		Fc	0
concretament	O	D4	concretament	O -1.0	rg	concretament	32	ADJUNCT		RG	0
a	O	P	a	O -1.0	sp	a	25	CONJUNCT		SPS00	0
Ca	B-LOC	N5	ca	B-LOC 0.94	np	ca	32	SN		NP00000	0
n'	I-LOC	VJ	n'	I-LOC 0.91	pp	en	33	SN		PP3CN00	0
Amat	I-LOC	VC	amar	I-LOC 0.91	np	amat	34	SN		NP00000	0

### Bibliografia:

- M. Ciaramita and Y. Altun. 2006. Broad-coverage sense disambiguation and information extraction with super-sense sequence tagger. In *Proceedings of the EMNLP*.  
G. Attardi, F. Dell'Orletta, M. Simi, A. Chanev, and M. Ciaramita. 2007. Multilingual dependency parsing and domain adaptation using DeSR. In *Proceedings the CoNLL Shared Task Session of EMNLP-CoNLL 2007*.  
Jordi Atserias and Bernardino Casas and Elisabet Comelles and Meritxell González and Lluís Padró and Muntsa Padró. FreeLing 1.3: Syntactic and semantic services in an open-source NLP library. *Proceedings of the fifth international conference on Language Resources and Evaluation (LREC 2006)*, ELRA. Genova, Italy. May, 2006.