

T A L N

Tractament Automàtic del Llenguatge Natural

Universitat Pompeu Fabra – Fundació Barcelona Media

<http://www.recerca.upf.edu/taln>



Jornada del Processament Computacional del Català

26 DE MARÇ DE 2009, PALAU ROBERT (BARCELONA)

Qui Som?

- *Investigadors senior*

Nadjet Bouayad-Agha (Prof. Assoc. UPF, BM)

Leo Wanner (ICREA i UPF, Cap de grup)

- *Investigadors*

Gerard Casamayor (BM, TICMA Master Program)

Simon Mille (BM)

- *Estudiants de doctorat*

Gabriela Ferraro (DTIC)

Vanesa Vidal (IULA - BM)



Línies de recerca

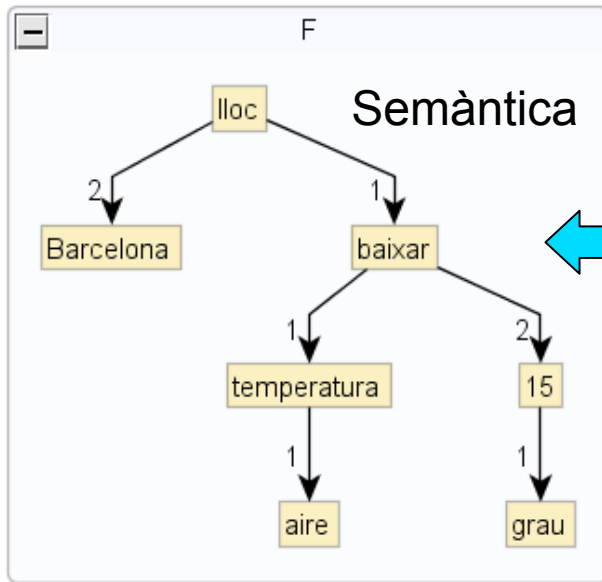
El treball del TALN es centra en alguns camps específics del processament automàtic del llenguatge natural

- generació multilingüe de llenguatge natural i altres continguts
- reescriptura: resum automàtic, paràfrasi, traducció, etc.
- lexicologia computacional: diccionaris de règims i de col·locacions
- aprenentatge automàtic orientat a l'adquisició de recursos lingüístics

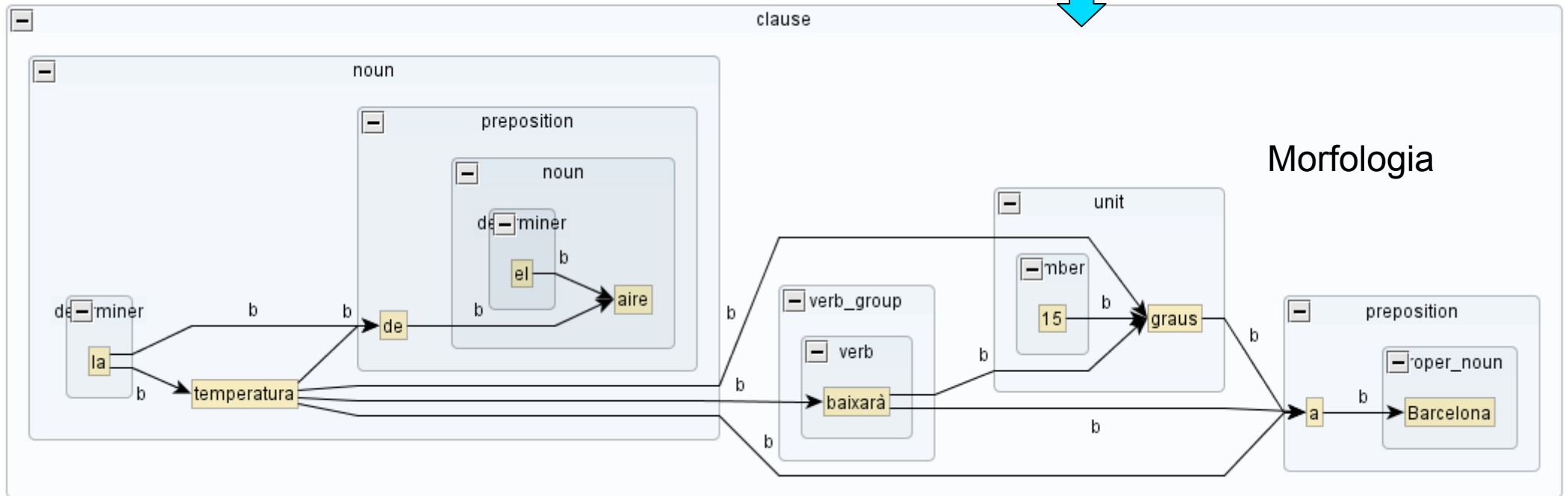
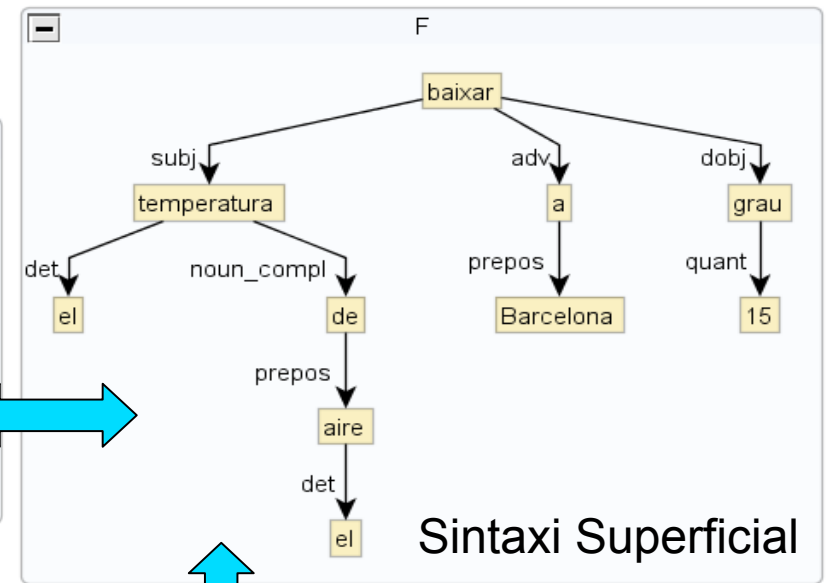
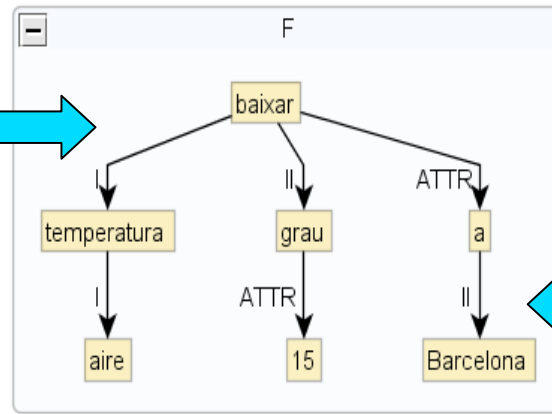
El marc lingüístic del nostre treball es la Teoria Sentit-Text (TST),

- Dependències
- Multinivell (mapeig entre nivells es duu a terme mitjançant gramàtiques i diccionaris)
- Holístic (cobreix tots els nivells i dimensions d'un model lingüístic)





Sintaxi Profunda



La temperatura de l'aire baixarà 15 graus a Barcelona.

Projectes (I)

Projectes finalitzats:

- Generació Multimodal i Multilingüe de □ butlletins sobre la qualitat de l'aire (MARQUIS, EDC-11258)
- Paràfrasi i resum multilingüe de documents de patents (PATExpert, FP 6-028116)

Projectes en curs:

- i3media (Generació de comentaris de partits de fútbol, generació d'instruccions d'entrenament i contribució a la conversa d'avatars)
- □ Col·locacions i tecnologies lingüístiques: cap a un entorn d'aprenentatge basat en la Web (COLOCATE, MCI FFI2008-06479-C02-02/FILO)

Projecte intern: Etiquetat d'un corpus de dependències sintàctiques per al castellà



Projectes (II)

- **Generació Multimodal i Multilingüe de □ butlletins sobre la qualitat de l'aire (MARQUIS, EDC-11258)**

Sèries de temps de □ pol·lució de cinc regions d'Europa

⇒ Generació en vuit llengües: Alemany, Anglès, Castellà, **Català**, Finès, Francès, Polonès i Portuguès

1	1	11	102	21	210	31	1011	41	1112
2	2	12	110	22	211	32	1012	42	1120
3	10	13	111	23	212	33	1020	43	1121
4	11	14	112	24	220	34	1021	44	1122
5	12	15	120	25	221	35	1022	45	1200
6	20	16	121	26	222	36	1100	46	1201
7	21	17	122	27	1000	37	1101	47	1202
8	22	18	200	28	1001	38	1102	48	1210
9	100	19	201	29	1002	39	1110	49	1211
10	101	20	202	30	1010	40	1111	50	1212



Aquest matí, la concentració d'ozó va baixar 40mg/m3 a Barcelona.

El nivell d'alerta va ocórrer ahir a les 15.00hs.

A Hèlsinki la concentració de PM10 ha estat baixa durant tota la setmana.



Eines i recursos pel català

→ *aim unes gramàtiques per a generar text a partir de representacions abstractes, en el domini del medi ambient (inclús oracions relatives, subordinades, coordinacions, pasives, etc.)*

→ *aim els recursos lèxics necessaris per la generació*

- *diccionari de mapeig entre unitats semàntiques i lèxiques, 300 unitats semàntiques*
- *diccionari de règim, 500 entrades*
- *TLM (Morfologia de doble nivell)*



Treball futur per el català

- *Etiquetatge d'un corpus de català de dependències de diversos nivells (sintaxi de superfície, sintaxi profunda i semàntica)*
- *Desenvolupament d'un parser de sintaxi de dependències*
- *Ampliació de la cobertura de las gramàtiques de generació*
- *Adquisició de recursos lèxics*

