

El corpus de l'IULA: explotació 1

El corpus de l'IULA: explotació

Seminari
IULA, 5 de març de 2007

Carme Bach
carme.bach@upf.edu

Institut Universitari de Lingüística Aplicada (UPF)

Barcelona, 5 de març de 2007

El corpus de l'IULA: explotació 2

Guió de la sessió

1. Corpus i subcorpus
2. Eines per a l'explotació
3. Explotació del Corpus de l'IULA

Institut Universitari de Lingüística Aplicada (UPF)

Barcelona, 5 de març de 2007

El corpus de l'IULA: explotació 3

Definició de *corpus*

- Conjunt organitzat de textos, emmagatzemats en suport informàtic de què és possible extreure informació.

Institut Universitari de Lingüística Aplicada (UPF)

Barcelona, 5 de març de 2007

El corpus de l'IULA: explotació 4

Ús de corpus: Les **W** de la recerca

- [Why?](#)
- [What/Which?](#)
- [Where?](#)
- [Who?](#)
- [How?](#)
- [When?](#)

Institut Universitari de Lingüística Aplicada (UPF)

Barcelona, 5 de març de 2007

El corpus de l'IULA: explotació 5

Ús de corpus en la recerca

- Why?
 - Validació d'hipòtesis
 - Documentació d'exemples
 - Documentació de contraexemples
 - Banc de proves experimental

Institut Universitari de Lingüística Aplicada (UPF)

Barcelona, 5 de març de 2007

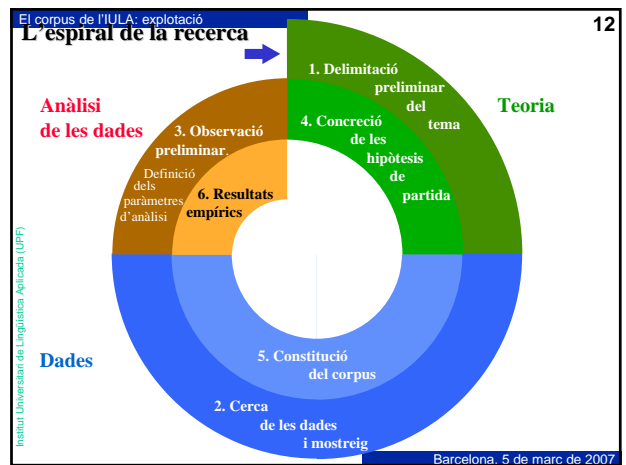
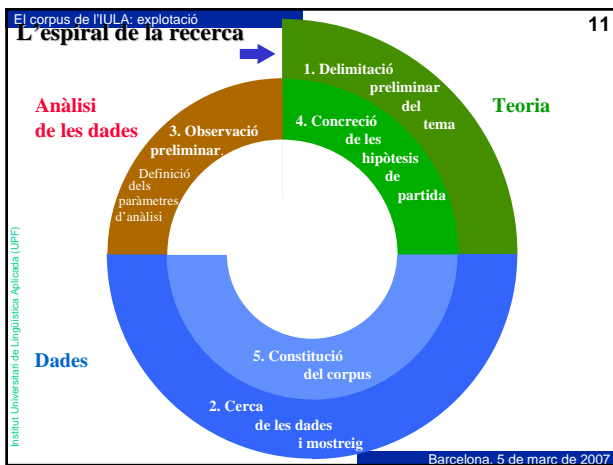
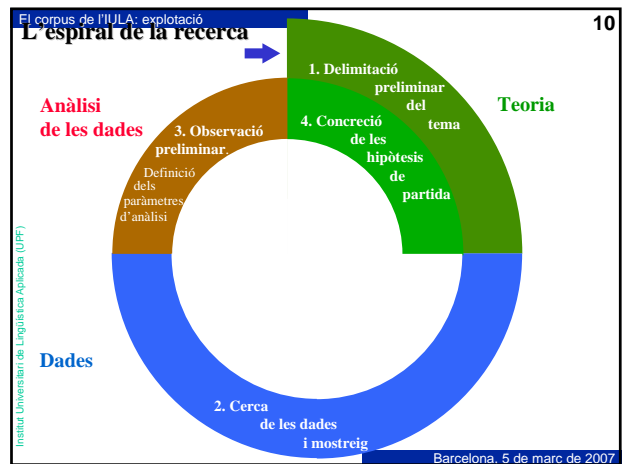
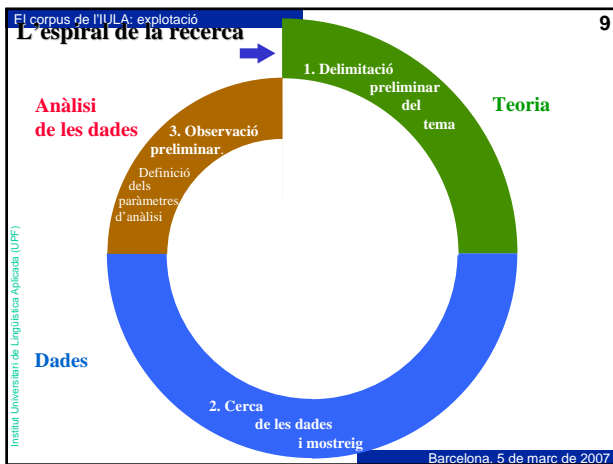
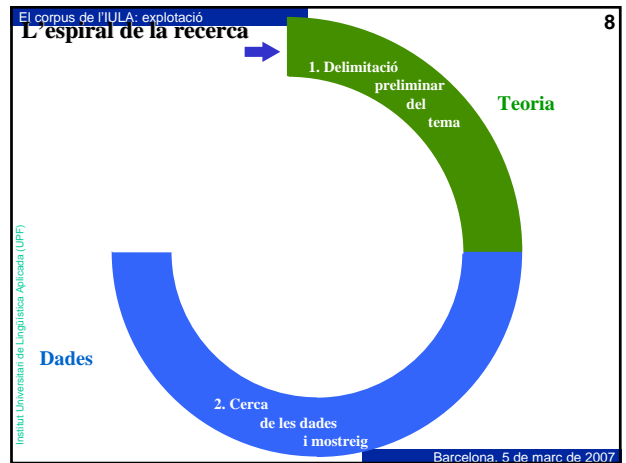
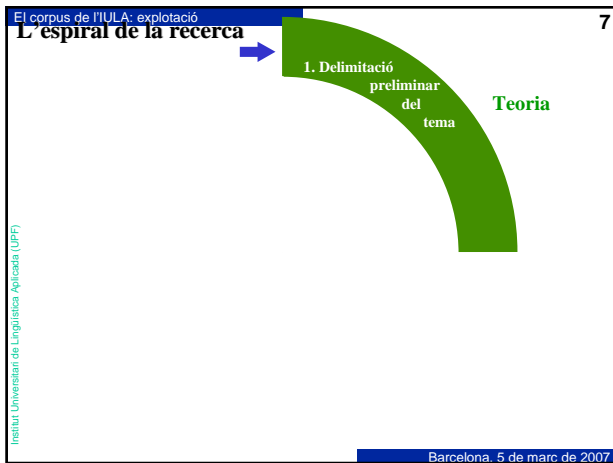
El corpus de l'IULA: explotació 6

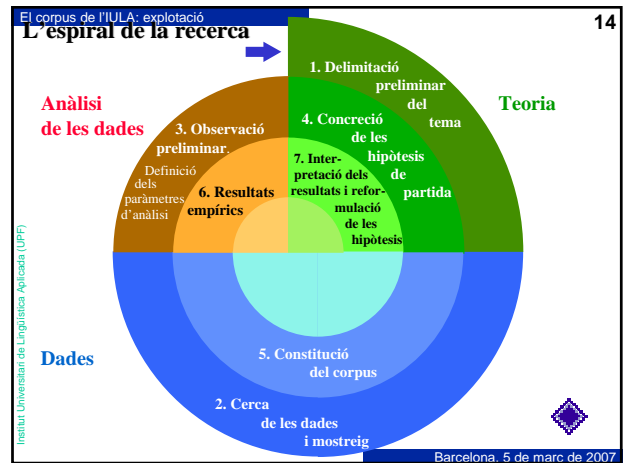
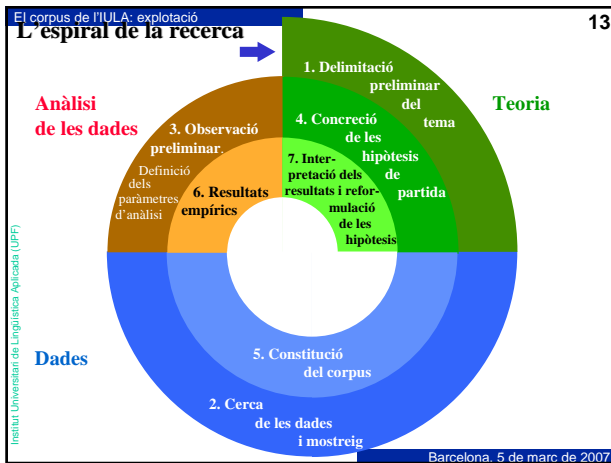
L'espiral de la recerca



Institut Universitari de Lingüística Aplicada (UPF)

Barcelona, 5 de març de 2007





- El corpus de l'IULA: explotació
- ### Ús de corpus: Orientació de la recerca
- What?
 - Teòrica
 - Descriptiva
 - Aplicada
 - Contrastiva
- Institut Universitari de Lingüística Aplicada (UFL)
- Barcelona, 5 de marc de 2007

- El corpus de l'IULA: explotació
- ### Què cerquem en els corpus?
- Informació sobre text / discurs
 - Informació sobre unitats
 - Quina mena d'informació?
 - Fenòmens lingüístics
 - Ocurrences concretes
- Institut Universitari de Lingüística Aplicada (UFL)
- Barcelona, 5 de marc de 2007

- El corpus de l'IULA: explotació
- ### Tipus de corpus
- Lexicogràfics
 - Lèxics
 - Textuals
 - escrits
 - orals
 - Altres
- Institut Universitari de Lingüística Aplicada (UFL)
- Barcelona, 5 de marc de 2007

- El corpus de l'IULA: explotació
- ### Selecció i delimitació
- Orientació de la recerca
 - Fenomen o objecte d'estudi
 - Accessibilitat de les fonts
 - Necessitat o no de tractament
 - Dimensions
- Institut Universitari de Lingüística Aplicada (UFL)
- Barcelona, 5 de marc de 2007

Ús de corpus en la recerca

- Where? On cercar els corpus?
 - Internet com a eina de cerca per aconseguir textos
 - Creació d'un nou corpus
 - Aprofitament de recursos ja existents



Ús de corpus en la recerca

- Who? (Corpus disponibles a l'IULA)
- CT
 - corpus genoma
 - corpus 92
 - corpus de premsa
 - corpus textual especialitzat
- Retoc (repertori electrònic de corpus orals catalans)
- CUVAL
 - corpus *La Canonja*
 - corpus *Codeswitching*
 - corpus narratives



Ús de corpus en la recerca

- How? De quina manera podem treballar els corpus?
 - Què hi busquem?
 - tractament textual (word)
 - informatització de les dades en format BD (accés)
 - explotació de les dades (informació lingüística / no)



Eines de Microsoft Windows

- Word: buscar i marcar, nombre de paraules, macros
- Access: control i gestió de bases de dades
- Excel: creació de gràfiques de barres, formatgets ...

Altres eines

- Editors de text: Editplus, Editpad
- Perl

Ús de corpus en la recerca

- When?
Quan vulgueu:
- [Corpus disponibles i eines de cerca de concordances](#)
- [El corpus de l'IULA](#)

Disponibilitat de corpus textuals complets

- [Biblioteca Virtual Miguel de Cervantes](#)
 - Textos generals i especialitzats classificats per matèries, en castellà.
- [Proyecto Gutenberg](#)
 - Textos generals i especialitzats en diverses llengües (EN, FR, ES, CA), alguns classificats per matèries.
- [BNC \(British National Corpora\)](#)

L'explotació de corpus

- Objectiu:
 - Dominar eines informàtiques d'anàlisi de textos que facilitin l'obtenció de dades:
 - Eines de cerca de concordances
 - Eines de consulta més dirigida (orientades cap a una finalitat determinada)

Eines de consulta de concordances

- [WebCorp](#)
 - Utilitza pàgines d'internet com a corpus per buscar concordances.
 - Llengua de les concordances, independent.
- [Corpus CREA](#)
 - Corpus de Referencia del Español Actual (RAE).
 - Permet restringir les cerques temàticament.
- [Simple Search of BNC-world](#)
 - Permet fer consultes reduïdes sobre el BNC.
- [PDL \(portal de dades lingüístiques\)](#)
 - Corpus de l'IEC, català.

El Corpus de l'IULA

- [Què conté?](#)
- [Eines d'explotació](#)

Dades del [corpus tècnic de l'IULA](#)

- Mòdul de text especialitzat:

Àrea	Total de Paraules	Idioma	Total de Paraules
• Dret	4.040.000	• Català	8.000.000
• Economia	3.250.000	• Castellà	9.160.000
• Medi Ambient	3.840.000	• Anglès	3.050.000
• Informàtica	2.500.000	• Francès	570.000
• Medicina (Genoma 3.100.000)	7.900.000	• Alemany	750.000
TOTAL	21.530.000	TOTAL	21.530.000

- Mòdul de text general:

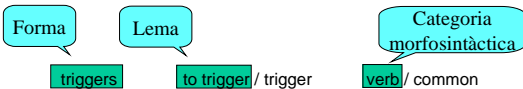
- Premsa
- Corpus 92
- D'altres

TOTAL 10.000.000 paraules

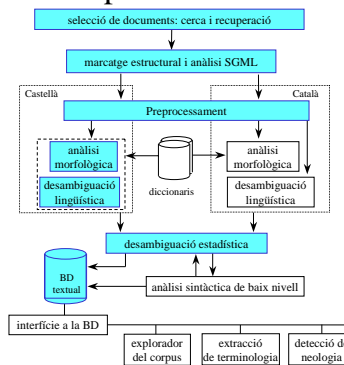
Característiques del corpus

- Etiquetatge morfològic.
- Lematització.
- Desambiguació lingüística i estadística.

signal from the receptor *triggers* binding of GTP to the RAS protein and GTP-RAS transmits the signal onwards in the cell



Cadena de processament del CT



El corpus de l'IULA: explotació 31

Explotació de corpus

- [Sistemes majoritàriament estadístics](#)
- [Sistemes majoritàriament lingüístics](#)


Institut Universitari de Lingüística Aplicada (UPF)

Barcelona, 5 de marc de 2007

El corpus de l'IULA: explotació 32

Explotació CT IULA amb una eina de base estadística

- [El projecte JAGUAR](#)
 - Projecte en proves. (Independent de llengua)
 - Per a més informació: jorge.vivaldi@upf.edu
 - Què fa?:
 - Freqüència.
 - Mesures sobre del grau d'associació entre paraules.
 - Mesures sobre la distribució de les paraules en el text.
 - Mesures de similitud.



Barcelona, 5 de marc de 2007

El corpus de l'IULA: explotació 33

Sistemes lingüístics

- [bwanaNet](#)
- Utilització d'informació lingüística per a l'explotació de corpus:
 - Patrons lingüístics
 - Anàlisi sintàctica

Institut Universitari de Lingüística Aplicada (UPF)

Barcelona, 5 de marc de 2007

El corpus de l'IULA: explotació 34

Què es bwanaNet? (1)

- Eina per a la interrogació del corpus tècnic de l'IULA, via Internet:
 - <http://bwananet.iula.upf.edu/>
- Eina lingüística amb informació estadística.

Institut Universitari de Lingüística Aplicada (UPF)

Barcelona, 5 de marc de 2007

El corpus de l'IULA: explotació 35

Què es bwanaNet? (2)

- Textos d'entrada marcats lingüísticament.
- Útil per a la cerca per patrons lingüístics.
- Combinació amb dades de freqüència.

Institut Universitari de Lingüística Aplicada (UPF)

Barcelona, 5 de marc de 2007

El corpus de l'IULA: explotació 36

Part pràctica

- Cerca en textos en castellà, català o anglès.
- Possibilitat de consultar textos paral·lels (que són els uns traduccions dels altres).

Institut Universitari de Lingüística Aplicada (UPF)

Barcelona, 5 de marc de 2007

El corpus de l'IULA: explotació 37

bwanaNet: extracció de concordances

Forma		
Lema	asma	
Etiqueta		JQ-66

⇒

asma bronquial
asma crònica
asma alèrgica

Forma		
Lema	economia	
Etiqueta		JQ-?? H??

⇒

economia agrària
economia pública
economia cerrada
economia cruzada
economies estocàstiques

Institut Universitari de Lingüística Aplicada (UPF)

Barcelona, 5 de marc de 2007

El corpus de l'IULA: explotació 38

bwanaNet: extracció de concordances

Forma			
Lema	derecho		
Etiqueta	P	(A??)1	N5-?? (JQ-66)1

⇒

derecho de información general
derecho de examen
derecho del creditor
derechos del comprador
derecho de los poseedores anteriores

Forma			
Lema	presentar		
Etiqueta	V??????	^(Z)10	P

⇒

presentar a
presentada por
presenta gran analogía con
presentando escritura pública en

Institut Universitari de Lingüística Aplicada (UPF)

Barcelona, 5 de marc de 2007

El corpus de l'IULA: explotació 39

bwanaNet: extracció de concordances a partir de la cerca per formants cultes

Forma		
Lema	*itis	
Etiqueta	N5-66	JQ-66

⇒

rinitis perenne
rinoconjuntivitis estacional
bronquitis asmàtica
dermatitis atòpica
alveolitis alèrgica

Forma		
Lema	rino*hemo*	
Etiqueta	N5-66	

⇒

rinorrea
rinosinusitis
hemorragia
hemoptisis
hemoglobina

Institut Universitari de Lingüística Aplicada (UPF)

Barcelona, 5 de marc de 2007

El corpus de l'IULA: explotació 40

bwanaNet: càlcul de freqüències

Last= a:[word="genoma"] @[pos="JQ.*" :: ((a.doc_area2="o")) within text; group Last matchend lemma by match lemma;

genoma humano	200
nuclear	22
completo	21
mitocondrial	20
vírico	10
haploide	10
celular	10
bacteriano	7
entero	5
doble	4
eucariota	3
eucariótico	3

Institut Universitari de Lingüística Aplicada (UPF)

Barcelona, 5 de marc de 2007

El corpus de l'IULA: explotació 41

Aplicació de la consulta de corpus multilingües a la traducció (1)

- Freqüència d'aparició d'un terme respecte a un altre amb una altra llengua a traduir (temàtica especialitzada).
- Detecció de variants denominatives útils per a la traducció:

Genetic disorder:

- Afección genética
- Desorden genético
- **Transtorno genético**
- Anomalia genética
- Alteración genética

Institut Universitari de Lingüística Aplicada (UPF)

Barcelona, 5 de marc de 2007

El corpus de l'IULA: explotació 42

Aplicació de la consulta de corpus multilingües a la traducció (2)

- Preposicions que s'afegeixen a la traducció, cal adjuntar-hi també un determinant?
- DNA replication
 - Replicación de ADN
 - Replicación de DNA
 - **Replicación del DNA**
 - Replicación del ADN
 - Replicación

Institut Universitari de Lingüística Aplicada (UPF)

Barcelona, 5 de marc de 2007

Modes d'interrogació

- Concordances (KWIC)
 - Simple forma/lema
 - Estàndard seqüències de forma/lema/categoria
 - Complexa llenguatge CQP
- Freqüències
 - Parcial (sobre un grup de documents)
 - Sobre tot el CT

Seqüència d'interrogació

1. Selecció de l'idioma de l'eina.
2. Selecció de la llengua dels documents.
3. Selecció consulta monolingüe/multilingüe.
4. Selecció dels documents.
5. Selecció del tipus de consulta.
6. Definició de la consulta.
7. Visualització dels resultats.

Consulta monolingüe/multilingüe

- Consultes sobre corpus alineats a nivell de frase.
- Restriccions morfològiques als documents paral·lels que s'apliquen a nivell de frase.

Selecció dels documents

- Selecció de dominis.
- Selecció de subdominis.
- Selecció per quantitat de paraules.
- Selecció de documents.
- Recuperació d'una selecció anterior.
- Selecció de tot el corpus.

Concordança simple

- Interrogació sobre forma/lema concret.
- Context complet/parcial.

Concordança estàndard

- Seqüències de formes/lemes/categories.
- Factor de repetició (aplicable a categories).
- Negació (aplicable a una categoria).
- Ancoratge de la concordança en relació a la unitat textual en què apareix. [head, s, p, list]
- Context total/parcial (ample que cal definir).
- Selecció del format dels resultats (lema/forma/categoria) i ordenació (lema/categoria).
- Demanda d'informació addicional (estatus del document, subdomini, tipus de document).
- Restriccions sobre la llengua dels documents paral·lels.

Concordança complexa (I)

- Llenguatge d'interrogació CQP.
- Màximes possibilitats d'interrogació.
 - Número il·limitat d'unitats.
 - Combinacions forma/lema/categoria.
 - Consultes multilingües.
 - Freqüències sobre formes, lemes, categories.
- Ajuda disponible en línia.

Concordança complexa (II)

- Accessos a la informació per parells de valors:
 - Atribut – Valor
 - Atributs del CT del IULA:
 - Forma: *word* [word="exemples"]
 - Lema: *lemma* [lemma="exemple"]
 - Categoria gramatical: *pos* [pos="N.*"]
 - Combinació d'atributs mitjançant operacions lògiques: ("&" (and), "&!" (and not), "|" (or))
- Ús d'expressions regulars ("*", ".", ".")

Concordança complexa (III)

- Cerques detallades por categories:
 - [Etiquetaris de l'IULA](#)
 - [pos="V.1(S)P"]
- Agrupació:
 - [pos="N5.*"] [pos="JQ.*"] [pos="H.*"]
- Disjunció:
 - [lemma="superficie"] ([lemma="terrestre"] [word="de"]
 - [lemma="el"] [lemma="tierra"])

Concordança complexa (IV)

- Càlcul de freqüències sobre:
 - Paraules
 - [pos="V.*"] <- patró de cerca
 - group Last match lemma ; <- acció
 - Seqüències
 - [pos="N5.*"] [pos="JQ.*"]
 - group Last matchend lemma by match lemma;

Concordança complexa (V)

- Ordenació del resultat:
 - [pos=N5.*] [pos="P"] [pos="N5.*"] [pos="JQ.*"]
 - Atribut inicial:
 - sort Last by lemma on match;
 - Atribut final:
 - sort Last by lemma on matchend;
 - Un altre atribut (marcat a la query amb @):
 - sort Last by lemma on target;

Concordança complexa (VI)

- Previsualització
 - Ordenació.
 - Amplada del context.
 - Mostrar/ocultar anotacions estructurals.
- Visualització
 - cat Last;
 - cat Last 0 10;

Unitats fora de context

- Genera llista de formes/lemes/categories segons una freqüència mínima a especificar.
- Freqüència mínima per defecte: 4.
- Format de sortida: text o HTML.

Freqüències sobre tot el CT

- Llistat de freqüències de formes, lemes, categories.
- Llistat de freqüències sobre determinades seqüències (màxim 3) de formes, lemes, categories.
- Especificar els atributs (excepte lema).

Emmagatzemament dels resultats

- Concordança
 - Guardar en format HTML i després obrir amb word
 - Enllaç **Resultats en format text**
 - Limitació:
 - 2000 línies (UPF)
 - 50 línies (des de fora UPF)
- Freqüències
 - Resultat en format text.

El corpus de l'IULA: explotació

Seminari
IULA, 5 de març de 2007

Carne Bach
carne.bach@upf.edu