

OSLIN - CA

LÈXIC OBERT FLEXIONAT DE CATALÀ

OSLIN

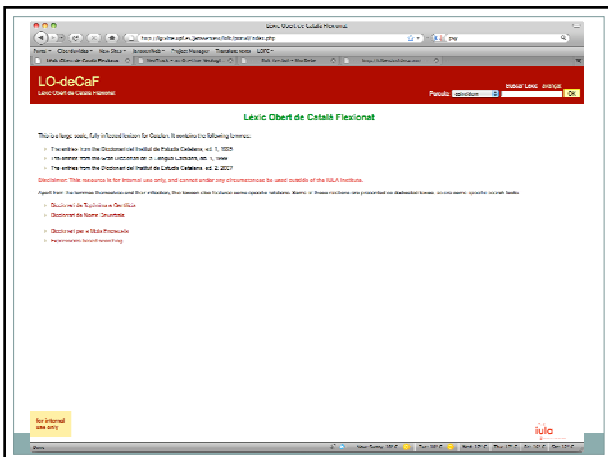
- **Open Source Lexical Information Network**
- **Language Independent Set-Up**
 - Originally built for Portuguese (MorDebe)
 - Intended to be built for many languages
- **Modular Design**
 - Built up from core data
 - Expandable with any type of data - relational
- **Multi-Purpose**
 - Linguistic and Computational Research
 - General Service

Lexicographic Control

- **High Quality Lexicon**
 - Manual Quality Control
 - Any error is one too many
- **Meta-Lexicographic**
 - Dictionaries are good language resources
 - Not complete (enough)
- **Constant Maintenance**
 - Eradicate last errors
 - Keep up-to-date
 - Keep from disappearing

Catalan Lexicon

- **Under development since November**
- **Contains major Catalan dictionaries**
 - DIEC1, GDLC, DIEC2
- **Currently 103.553 lemmas**
 - And 717.829 inflected forms
 - To be expanded with additional lemmas
- **Still a “confidential” project**
 - To be negotiated with publishers
- **Only basic information (for the moment)**



User Perspective

GENERAL LANGUAGE USER

General Service Web-Site

- **Portuguese OSLIN web-site**
 - MorDebe launched January 2005
 - Portal launched January 2007
- **Very popular resource**
 - Started with around 600 users per day
 - Currently topping at 7.000 users per day
 - Large language professionals
- **Does not exist for Catalan**
 - Would be very helpful in several ways
 - Less innovative than for Portuguese

Full-Form Lexicon

- **OSLIN contains full inflections**
 - Explicitly stored (not generated)
- **Consult inflection**
 - Verbal conjugations
 - Nominal conjugations
- **Reverse lookup**
 - No need to lemmatize first
 - Complicated with rule-based lexicon

Inherent Inflections

- **OSLIN models “extended paradigms”**
 - Including inflection-like derivations
 - Event nouns, property nouns, diminutives, adverbs, etc.
- **Presented with each lemma**
 - Presenting information in both directions
 - Also as dedicated dictionaries
- ***Diccionari de Noms Deverbals***
 - All dictionarized deverbal event nouns
 - Could/should be extended with all used forms
 - Not full *potential lexicon*

Source Tracking

- **OSLIN includes other resources**
 - Dictionaries DIEC1, DIEC2, GDLC
- **Explicitly keeps track of sources**
 - Each lemma contains information where it is found
- **Distinguish between status**
 - Words in dictionaries / privately added words
 - Exploit the status of dictionaries
- **Link to original sources**
 - Web-sites of GDLC and DIEC2
 - Indirectly provided semantic information

Gentilicis

- **OSLIN contains a database of Proper Names**
 - Including Toponyms
- **Explicitly links *gentilicis* to their toponym**
 - Represented with each gentileci
- **Dedicates *Diccionari de Gentilicis***
 - Information in the other direction
 - All dictionarized *gentilicis* (1.500 x2)
- **Future Project**
 - Make it (more) complete
 - Controversial but extremely useful

Fun Part

- **Word MasterMind – Català**
 - Guess a word of X letters
 - Not very scientific – but fun and promotional
- ***Diccionari per a Mots Encreuats***
 - Gives you all registered words matching a given pattern
 - Great for crossword puzzles
 - And for word mastermind
 - *Possibly for Scrabble™ as well*

Conclusion

- **Very useful resource for Catalan population**
 - We will attempt to provide it as a service
- **Relies on “go-ahead” from publishers**
 - Not illegal, but very disruptive
- **Will require maintenance**
 - And hence funding
 - And people with lexicographic capabilities
- **Will lead to questions**
 - Which should ideally be answered

Academic Perspective

RESEARCH TOOL

Morphology / Phonetics

- **Search for parts of words**
 - Both in lemmas and in word-forms
- **Dedicated search tool**
 - “Regular expression search”
 - Explicit endings and beginnings
 - Search for letters, vowel, or consonants
- **Shown to be useable**
 - Used by linguistic community in Portugal
 - Often used for classes and research

Natural Language Processing

- **OSLIN uses a proprietary format**
 - Meant to build resources, not discuss them
- **Several Export Functions**
 - Simple list of words
 - TreeTagger format (with IULA tagset)