# Words & Works


German Rigau i Claramunt

german.rigau@ehu.es

IXA group

Departamento de Lenguajes y Sistemas Informáticos

UPV/EHU

# Introduction

- Which Knowledge is needed by a concrete NLP system?
- Where is this Knowledge located?
- Which automatic procedures can be applied?

# Introduction

- Which Knowledge is needed by a concrete NLP system?
  - Phonology: phonemes, stress, etc.
  - Morphology: POS, etc.
  - Syntactic: category, subcat., etc.
  - Semantic: class, SRs, etc.
  - Pragmatic: usage, registers, TDs, etc.
  - Translations: translation links

# Introduction

- Where is this Knowledge located?
  - Human brain
  - Structured Lexical Resources:
    - Monolingual and bilingual MRDs
    - Thesauri
  - Unstructured Lexical Resources:
    - Monolingual and bilingual Corpora
  - Mixing resources

# Introduction

- Which automatic procedures can be applied?
  - Prescriptive approach
    - Machine-aided manual construction
  - Descriptive approach
    - Automatic acquisition from pre-existing Lexical Resources
  - Mixed approach

# Where is this Knowledge located?

- Human brain:
  - WordNet (Miller et al. 90)
    - Semantic Information v1.6 with 99,642 synsets
  - Comlex (Grishman et al. 94)
    - Syntactic information 38,000 English words
  - CYC Ontology (Lenat 95, Malesh et al. 96)
    - 900 person-year of effort to produce 100,000 terms
  - Mikrokosmos (Viegas et al. 98)
    - Ontology for MT with 5,000 concepts
  - SUMO (Niles & Pease 01)
    - IEEE ontology

# Where is this Knowledge located?

- Structured Lexical Resources (1)
  - Monolingual MRDs:
    - LDOCE
      - learner's dictionary
      - 35,956 entries and 76,059 definitions
      - 86% semantic and 44% pragmatic codes
      - controlled vocabulary of 2,000 words
      - (Boguraev & Briscoe 89)
      - (Vossen & Serail 90)
      - (Bruce & Guthrie 92), (Wilks et al. 93)
      - (Dolan et al. 93), (Richardson 97)

# Where is this Knowledge located?

- Structured Lexical Resources (2)
  - Other Monolingual MRDs:
    - Webster's (Jensen & Ravin 87)
    - LPPL (Artola 93)
    - DGILE (Castellón 93), (Taulé 95), (Rigau 98)
    - CIDE (Harley & Glennon 97)
    - AHD (Richardson 97)
    - WordNet (Harabagiu 98)
  - Bilingual MRDs
    - Collins Spanish/English (Knigth & Luk 94)
    - Vox/Harrap's Spanish/English (Rigau 98)

# Where is this Knowledge located?

- Structured Lexical Resources
  - Thesauri:
    - Roget's Thesaurus
      - 60,071 words in 1,000 categories
      - (Yarowsky 92), (Grefenstette 93), (Resnik 95)
    - Roget's II and The New Collins Thesaurus
      - (Byrd 89)
    - Macquarie's thesaurus
      - (Grefenstette 93)
    - Bunrui Goi Hyou Japanese thesaurus
      - (Utsuro et al. 93)

# Where is this Knowledge located?

- Structured Lexical Resources
  - Encyclopaedia
    - Grolier’s Encyclopaedia (Yarowsky 92)
    - Encarta (Richardson et al. 98)
    - Wikipedia (Horacio Tutorial!)
  - Others
    - Telephonic Guides
  - Mixing structured lexical resources
    - Roget’s Thesaurus and Grolier’s (Yarowsky 92)
    - LDOCE, WN, Collins, ONTOS, UM (Knight & Luk 94)
    - Japanese MRD to WN (Okumura & Hovy 94)
    - LLOCE, LDOCE (Chen & Chang 98)

# Where is this Knowledge located?

- Unstructured Lexical Resources
  - Corpora:
    - Proper Nouns (Hearst & Schütze 95)
    - Idiosyncratic Collocations (Church et al. 91)
    - Preposition preferences (Resnik and Hearst 93)
    - Subcategorization structures (Briscoe and Carroll 97)
    - Selectional preferences (Resnik 93; Ribas 95; McCarthy 01)
    - Thematic structure (Basili et al. 92)
    - Word semantic classes (Dagan et al. 94; Lin & Pantel 98)
    - Bilingual Lexicons for MT (Fung 95)
    - Semantic relations (Pennachiotti & Pantel 05)
    - Topic Signatures (Agirre et al. 04)

# Where is this Knowledge located?

- Using both structured and non-structured Lexical Resources
  - MRDs and Corpora
    - (Liddy & Paik 92)
    - (Klavans & Tzoukermann 96)
  - WordNet and Corpora
    - (Resnik 93), (Ribas 95), (Li & Abe 95), (McCarthy 01),
    - (McCarthy et al. 04)
    - (Agirre et al. 04)
    - (Cuadros et al. 04)

# International Projects on Knowledge Acquisition

- Japanese Projects
  - EDR (Yokoi 95)
    - Nine years project oriented to MT
    - Bilingual Corpora with 250,000 words
    - Monolingual, bilingual and coocurrence dictionaries
    - 200,000 general vocabulary
    - 100,000 technical terminology
    - 400,000 concepts

# International Projects on Knowledge Acquisition

- American Projects
  - WordNet (Miller 90)
    - Semantic Information
    - more than 123,000 words organised in 99,000 synsets
    - more than 116,000 relations between synsets
  - Comlex (Grishman et al. 94)
    - Syntactic information for 38,000 words
  - Cyc (Lenat 95)
    - common-sense knowledge
    - 100,000 concepts and 1,000,000 axioms
  - Pangloss (Knight & Luk 94) Omega (Hovy et al. 03)
    - PUM, ONTOS, LDOCE semantic categories, WordNet
  - Open Mind
    - http://www.openmind.org/

# International Projects on knowledge Acquisition

- European Projects
  - Acquilex I and II
    - LA from monolingual and bilingual MRDs and corpora
  - LE-Parole
    - Large-scale harmonised set of corpora and lexicons for all the EU languages
  - EuroWordNet
    - Multilingual WordNet for several European Languages
  - Meaning
    - Large-scale of LK from the web
    - Large-scale WSD

# Acquisition of knowledge from MRDs

- Syntactic Disambiguation (Dolan et al. 93)
- Semantic Processing (Vanderwende 95)
- WSD (Lesk 86), (Wilks & Stevenson 97), (Rigau 98)
- IR (Krovetz & Croft 92)
- MT (Knight and Luk 94), (Tanaka & Umemura 94)
- Semantically enriching MRDs
  - (Yarowsky 92), (Knight 93), (Chen & Chan 98), (Castillo et al. 03)
- Building LKBs
  - (Bruce & Guthrie 92)
  - (Dolan et al. 93)
  - (Artola 93), (Castellón 93), (Taulé 95), (Rigau 98)
  - (Mihalcea & Moldovan 01), (Castillo et al. 04)

# Acquisition of knowledge from MRDs

- Why MRDs?

  The conventional dictionaries for human use usually "contain spelling, pronunciation, hyphenation, capitalization, usage notes for semantic domains, geographic regions, and propiety; ethimological, syntactic and semantic information about the most basic units of the language" (Amsler 81)

- Main problems
  - Conventional dictionaries are not systematic
  - Dictionaries are built for human use
  - Implicit Knowledge
    - words are described/translated in terms of words

# MRDs and Semantic Knowledge

| | |
|---|---|
| **jardín_1_1** | Terreno donde se cultivan plantas y **flores** ornamentales. |
| **florero_1_4** | Maceta con **flores**. |
| **ramo_1_3** | Conjunto natural o artificial de **flores**, ramas o hierbas. |
| **pétalo_1_1** | Hoja que forma la corola de la **flor**. |
| **tálamo_1_3** | Receptáculo de la **flor**. |
| **miel_1_1** | Substancia viscosa y muy dulce que elaboran las abejas, en una distensión del esófago, con el jugo de las **flores** y luego depositan en las celdillas de sus panales. |
| **florería_1_1** | Floristería; tienda o puesto donde se venden **flores**. |
| **florista_1_1** | Persona que tiene por oficio hacer o vender **flores**. |
| **camelia_1_1** | Arbusto cameliáceo de jardín, originario de Oriente, de hojas perennes y lustrosas, y **flores** grandes, blancas, rojas o rosadas (Camellia japonica). |
| **camelia_1_2** | **Flor** de este arbusto. |
| **rosa_1_1** | **Flor** del rosal. |

# Words & Works


German Rigau i Claramunt

german.rigau@ehu.es

IXA group

Departamento de Lenguajes y Sistemas Informáticos

UPV/EHU