Resources for Language Understanding

- General Lexicons
- Dictionaries
- Specialized Lexicons
- Ontologies
- Grammars
- Textual Corpora
- Internet as an information source

General Lexicons

- Word repositories
  - Lemmaries, formaries, lists of words, phrasal lexicons
- Knowledge on words
  - Phonology
  - Morphology: part of specch, agreement
  - Syntax: category, subcategorization, argument structure, valency

  co-occurrence patterns
  - Semantics: semantic class, selectional restrictions
  - Pragmatics: use, register, domain

Dictionaries

- MRDs (Machine Readable Dictionaires)
- Types: general, normative, learner, mono/bilingual
- Size, content, organization
    - entry, sense, relations,
- Lexical databases
    - e.g. Acquilex LDB
- Other sources: enciclopaedias, thesaurus
    - e.g. Wikipedia

Specialized Lexicons

- Onomasticae
- Terminoligical databases
- Gazetteers
- Dictionaries of locutions, idioms
- Wordnets
- Acronyms, idioms, jaergon
- Date, numbers, quantities+units, currencies

Example: Using Gazetteers in Q&A systems

- Multitext (U.Waterloo)
  - Clarke et al, 2001, 2002
    - Structured data
      - Biographies (25,000), Trivial Q&A (330,000), Country locations (800), acronyms (112,000), cities (21,000), animals (500), previous TREC Q&A (1393), ...
    - 1 Tb of Web data
    - Altavista
- AskMSR (Microsoft)
  - Brill, 2002

Representation

- General purpose databases
- Textual databases
- Lexical databases
- Object oriented formalisms
- Object oriented databases
- Frames- hierarchical
- Unification-based formalisms

Ontologies

- Lexical vs conceptual ontologies
- General vs domain restricted ontologies
- Task ontologies, meta-ontologies
- Content, granularity, relations
- Interlinguas: KIF, PIF
- CYC, Frame-Ontology, WordNet, EuroWordNet, GUM, MikroKosmos
- Protegeé

WordNet I

- A large lexical database of English
- Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets)
- Synsets are interlinked by means of conceptual-semantic and lexical relations
-
- It interlinks specific senses of words
- It labels the semantic relations among words

http://wordnet.princeton.edu/

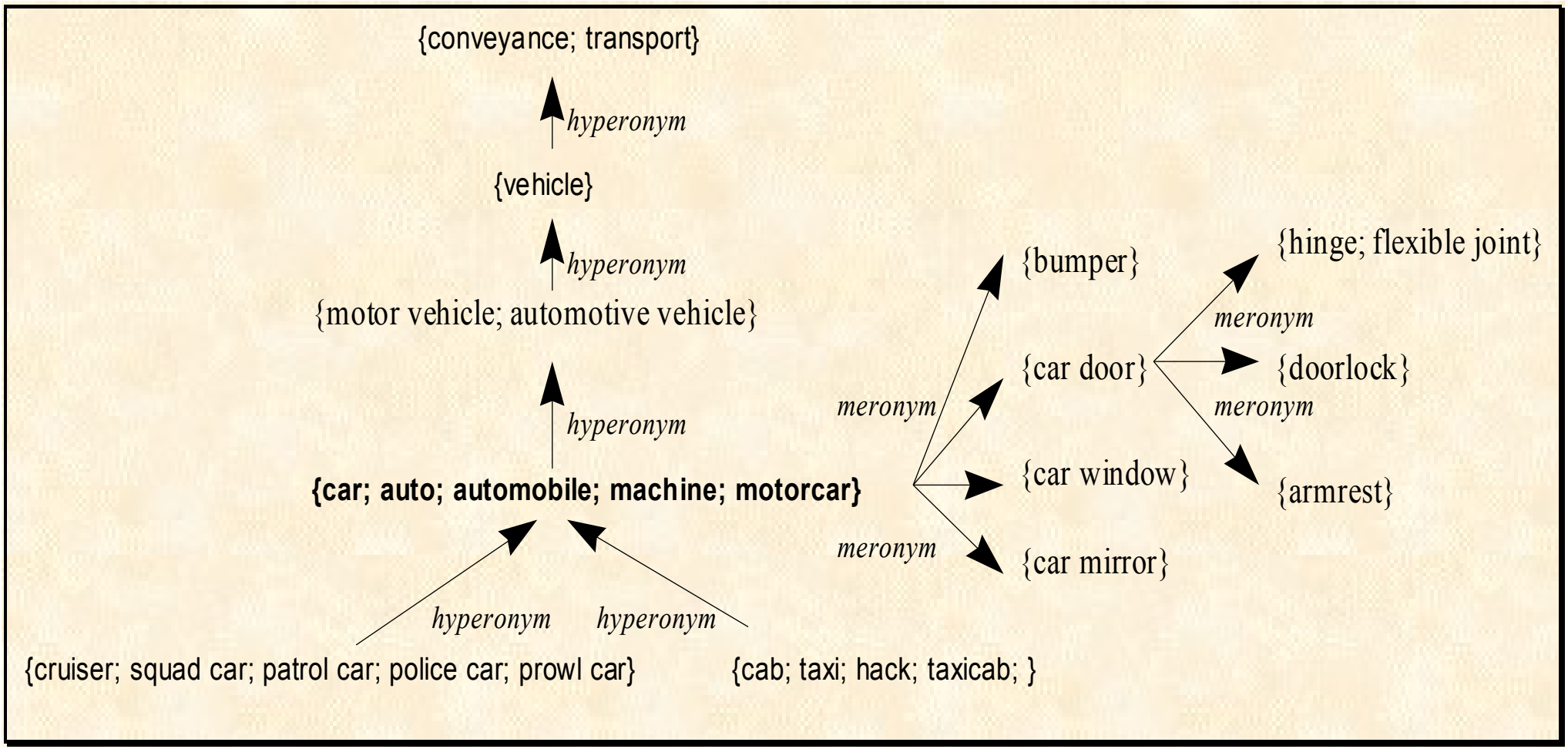
WordNet II

- Synonyms are grouped into unordered sets: **synsets**
- There are 117 000 synsets
- Word forms with several distinct meanings are represented in as many distinct synsets
- Each synset is linked to other synsets by means of a small number of conceptual relations
- Nouns, verbs and adjectives are represented in separated sets
- Nouns and verbs are represented in hierarchies
- Most frequent relation between nouns and verbs is the relation **hyperonymy** (inverse to **hyponymy** for nouns and **troponym** for verbs)

WordNet III

- Types (comomon nouns) and instances (specific names) are distinguished
- Other relations between nouns are:
  - meronym (part-whole)
  - has-member (member of)
  - antonym
- relations between nouns are:

Fragment of WN1.5

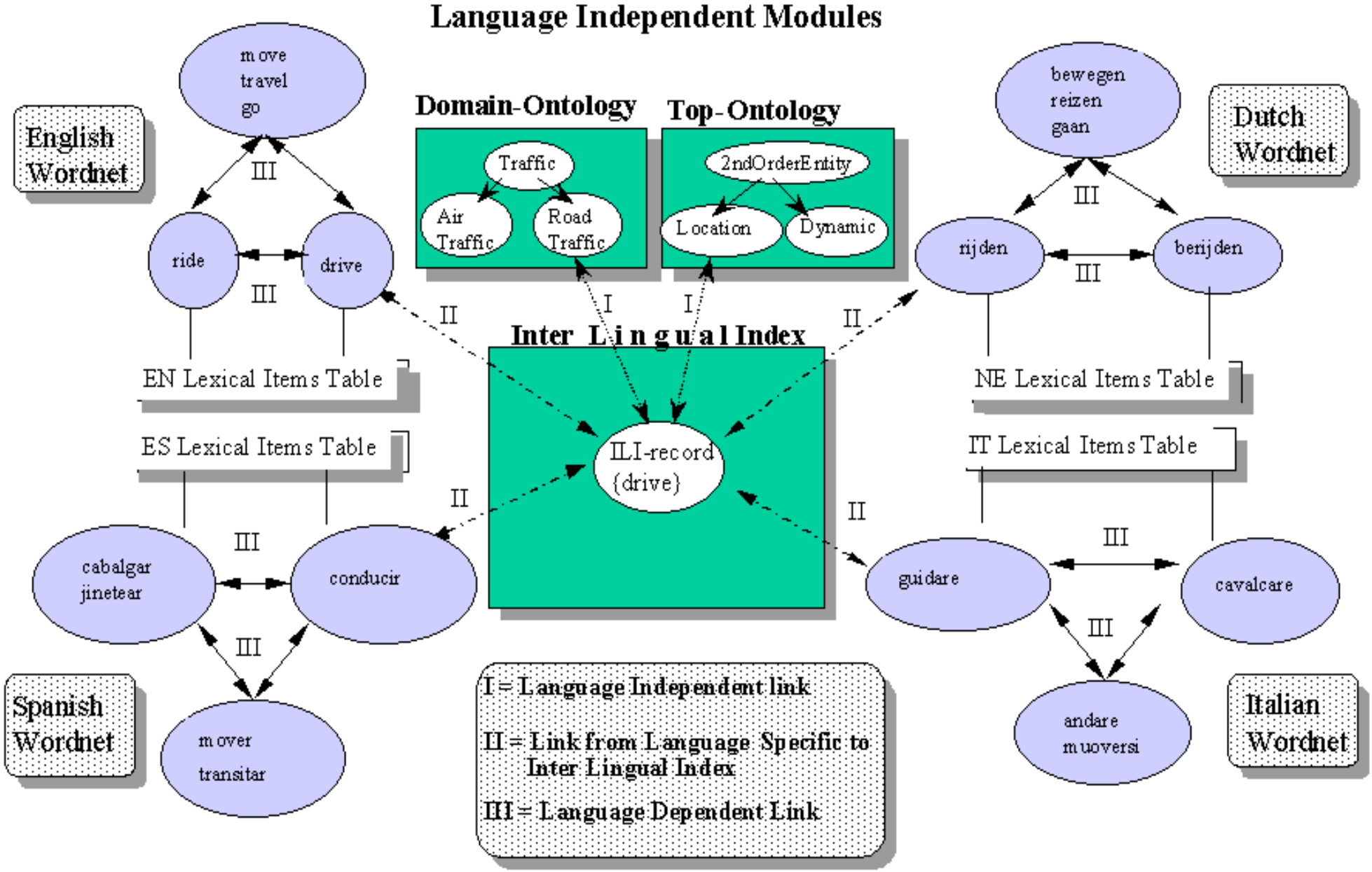


EuroWordNet

- Project LE-2 4003 Telematics Application Programme of the European Community
- Semantic networks in different languages (Integrated)
  - English Universidad de Sheffield
  - Dutch Univ. de Amsterdam
  - Italian I.L.C. de Pisa
  - Spanish UB, UPC, U.N.E.D
- Covers basically nouns and verbs (50.000 meanings for each language)
- Rich in semantic relationships
  - inter/intra lingual, inter/intra category
- EWN2
  - German, Czech, Estonian, French
- Extensions to Catalan, Galician and Basque
- Improvements

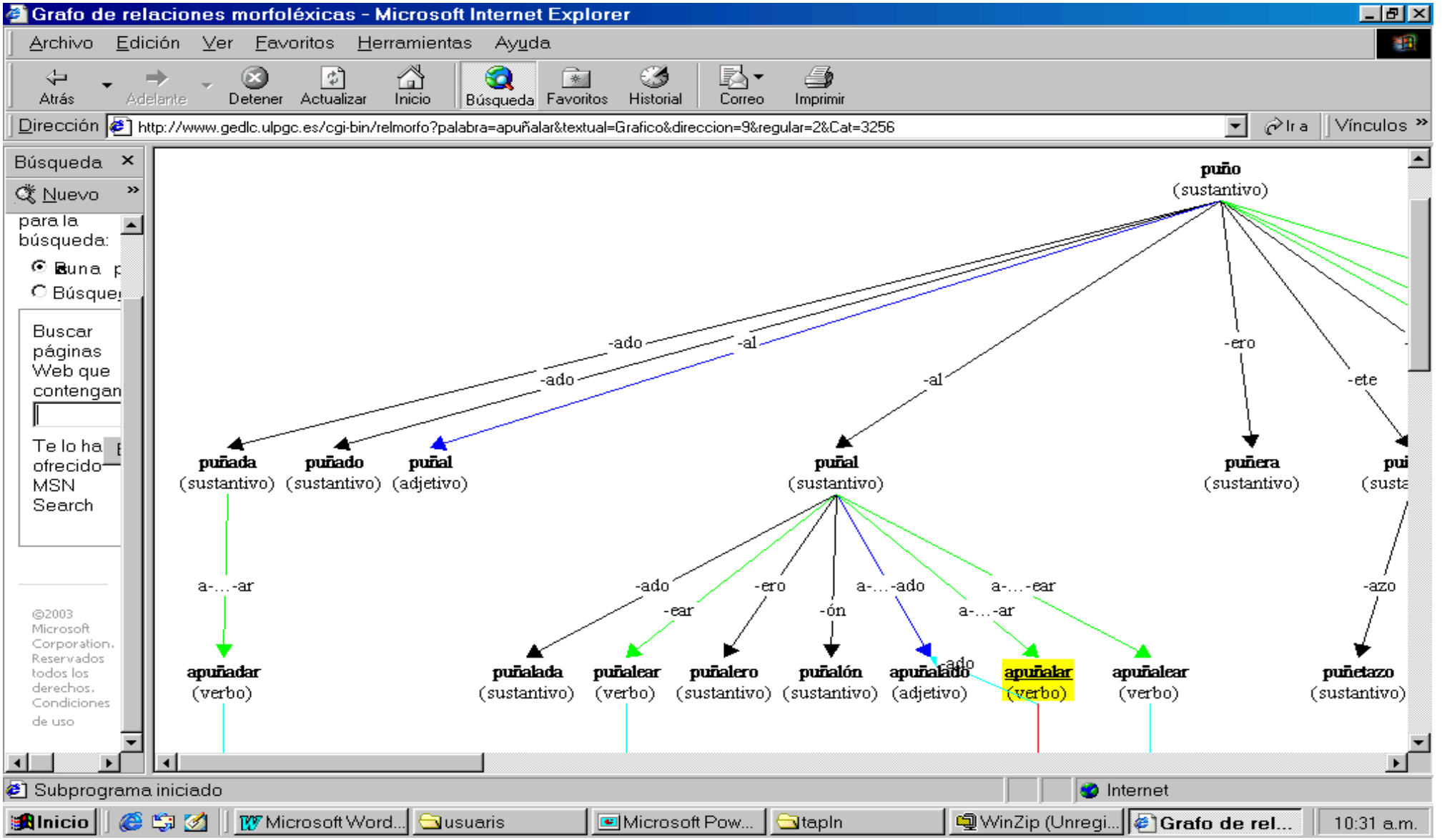
http://www.hum.uva.nl/~ewn/
http://www.lsi.upc.es/~nlp/

Architecture of the EuroWordNet Data Structure



Morpholexical Relations

U. Las Palmas (Santana)



Lexical information acquisition

- Dictionaries
  - Predefined internal structure
  - Some degree of coding in some contents
  - Internal relations (synonimy, hyponymy, ...)
  - (sometimes) restricted vocabulary
  - Some sistematics on building definitions

Grammars

- Morphological Grammars
- Syntactic Grammars
  - constituents
  - dependency
  - case
  - transformational
  - systemic
- Phrase-strucure vs Unification Grammars
- Probabilistic Grammars
- Coverage, language, tagsets

Textual Corpora

Information present in corpora

- Colocations
- Argument structure.
- Frecuency information
- Context
- Grammatical Induction
- Probabilistic Analysis.
- Lexical relations
- Examples of use.
- Selectional Restrictions
- Nominal compounds
- Idioms, ...

Corpus typology

- Raw corpus
- Horizontal or vertical Corpus
- Tagged corpora
- Parenthized corpora
- Treebanks

Raw Corpora

- Textual vs Speech
- Size (1Mw - 1Gw - 1TW)
- Few estructure (if any)
- Provide information not available in a more treatable way:
  - colocations, argumental structure, context of occurrence, grammatical induction, lexical relations, selectional restrictions, idioms, examples of use

Tagged Corpora

- Pos tagged (all tags are disambiguated)
- Lemma
- Sense (granularity of tagset, WN)
- Parenthised
  - parsed
- Parallel corpora
- Balanced, pyramidal, oportunistic corpora

Some examples of Corpora

- Brown Corpus
- ACL/DCI (Wall Street Journal, Hansard, ...)
- ACL/ECI (European Corpus Initiative)
- USA-LDC (Linguistic Data Consortium)
- LOB (ICAME, International Computer Archive of Modern English)
- BNC (British National Corpus)
- SEC (Lancaster Spoken English Corpus)
- Penn Treebank
- Susanne
- SemCor
- Trésor de la Langue Française (TLF)

Penn Treebank

- 1,3 millon wods, 40.000 sentences
- Wall Street Journal and other sources
- POS tagged
- Syntactically Parsed