# Basic NLP Tools

German Rigau i Claramunt
german.rigau@ehu.es
IXA group
Departamento de Lenguajes y Sistemas Informáticos
UPV/EHU

# Content

- Tools and Applications
  - Introduction
  - Basic Tools
    - Basic processing (Unix for Poets)
      - Tokenization, Sentence Splitting, Language Identifiers, ..
    - Stemming, lemmatization, POS tagging, ...
    - Named Entity Recognizers and Categorizers (NERC)
    - Parsing
    - Word Sense Disambiguation (WSD)
    - Coreference resolution: anaphoric references, ...
    - Semantic Role Labelling (SRL)
    - ...
    - Complete NLP suites

# Basic NLP Tools
# **Introduction**

- Public Catalogues
  - http://sinai.ujaen.es/timm/wiki/index.php/Recursos
  - http://ixa2.si.ehu.es/know2/index.php/Inventario_recursos
  - http://aclweb.org/aclwiki
  - ...
- NewsReader Deliverable D4.1
  - http://www.newsreader-project.eu/files/2012/12/NewsReader-316404-D4.1.pdf

# Morphological Analysis

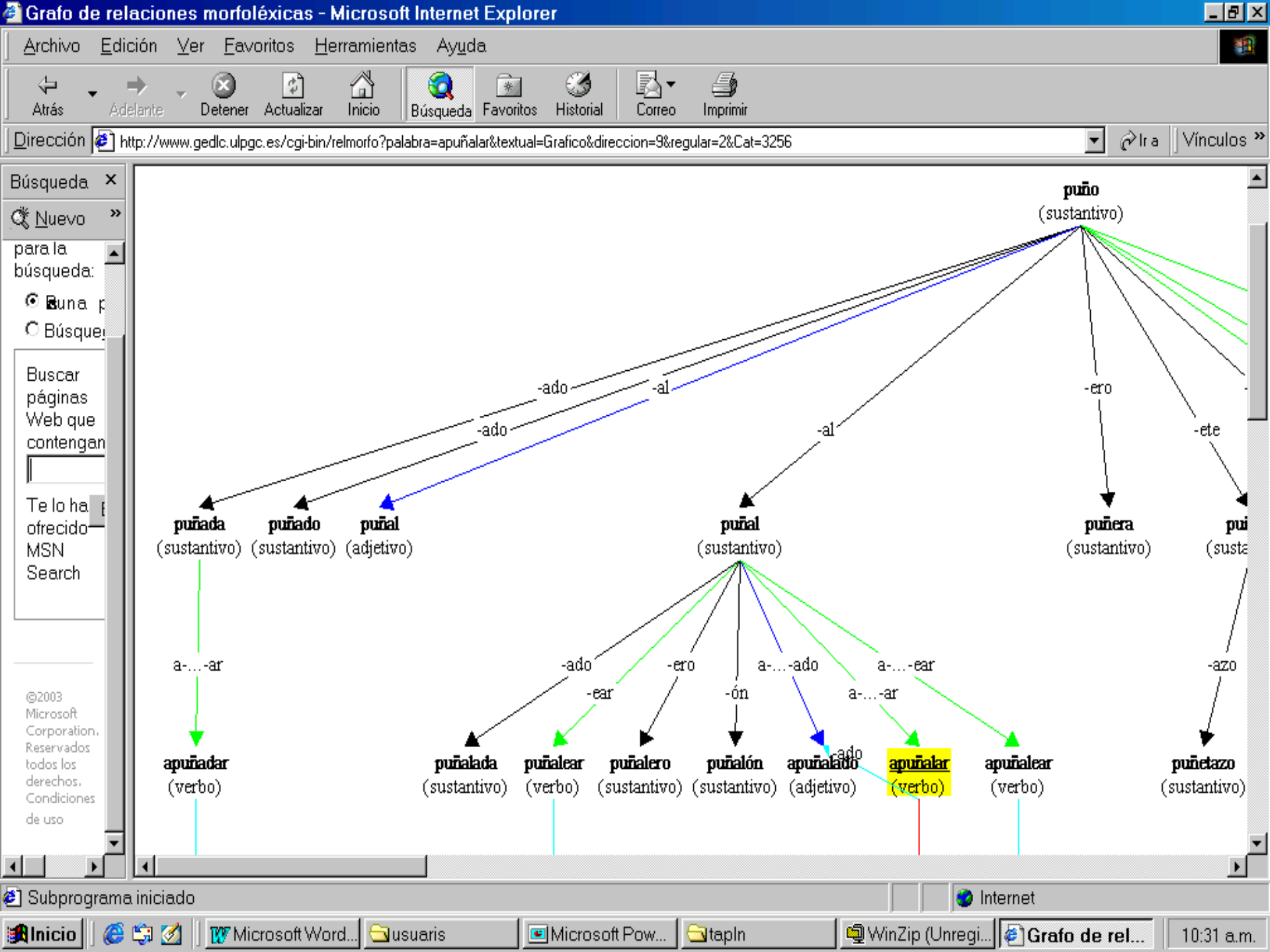
- Setting
- Systems
  - Morpholexical relationships (Octavio Santana)
  - Freeling (Lluís Padró)
  - IXA-pipeline
  - English stemmers
  - ...

# Morphological Analysis

- Morphology deals with the orthographic form of the words
- Morphological processes
  - Inflection: prefixes + root + suffixes (root, lemma, form)
  - Derivation: change of category
- Multi-word expressions: compounds, idioms, phrasal verbs, ...
- Grammatical categories, parts-of-speech
  - Open categories and closed (functional) categories
  - Lexicon
  - POS tags

# Morphological Analysis

- Main Parts-of-Speech
  - Open class words
    - Noun: common noun, proper noun (gender, number, ...)
    - Adjective: attributive, comparative ...
    - Verb: (number, person, mode, tense), auxiliary verbs
    - Adverb: place, time, manner, degree, ...
  - Closed class words
    - Pronoun: nominative, accusative, ... (anaphora)
    - Determiner: articles, demonstratives, quantifiers ...
    - Preposition:
    - Conjunction:


Grafo de relaciones morfoléxicas - Microsoft Internet Explorer
Archivo Edición Ver Favoritos Herramientas Ayuda
Atrás Adelante Detener Actualizar Inicio Búsqueda Favoritos Historial Correo Imprimir
Dirección http://www.gedlc.ulpgc.es/cgi-bin/relmorfo?palabra=apuñalar&textual=Grafico&direccion=9&regular=2&Cat=3256
Ir a Vínculos
Búsqueda
Nuevo
para la búsqueda:
Buscar páginas Web que contengan
Te lo ha ofrecido MSN Search
©2003 Microsoft Corporation. Reservados todos los derechos. Condiciones de uso
puño (sustantivo)
-ado
-al
-ado
-al
-ero
-ete
puñada (sustantivo)
puñado (sustantivo)
puñal (adjetivo)
puñal (sustantivo)
puñera (sustantivo)
a-…-ar
-ado
-ear
-ero
-ón
a-…-ado
a-…-ear
a-…-ar
-azo
apuñadar (verbo)
puñalada (sustantivo)
puñalear (verbo)
puñalero (sustantivo)
puñalón (sustantivo)
apuñalado (adjetivo)
-ado
apuñalar (verbo)
apuñalear (verbo)
puñetazo (sustantivo)
Subprograma iniciado
Internet
Inicio
Microsoft Word...
usuaris
Microsoft Pow...
tapln
WinZip (Unregi...
Grafo de rel...
10:31 a.m.

**Write your sentences**

Detenido en Barcelona el presunto jefe de las dos células islamistas desarticuladas

**Analysis options**

- [x] Multiword detection
- [x] Number recognition
- [x] Date/Time recognition
- [x] Named Entity detection
- [x] Quantities, ratios, and percentages

**Select language** Spanish

**Select output** PoS Tagging

Submit

## Analysis Results

| | |
|---|---|
| **Detenido** | *detener* VMP00SM |
| **en** | *en* SPS00 |
| **Barcelona** | *Barcelona* NP00000 |
| **el** | *el* DA0MS0 |
| **presunto** | *presunto* AQ0MS0 |
| **jefe** | *jefe* NCMS000 |
| **de** | *de* SPS00 |
| **las** | *el* DA0FP0 |
| **dos** | *dos* DN0CP0 |
| **células** | *célula* NCFP000 |
| **islamistas** | *islamista* AQ0CP0 |
| **desarticuladas** | *desarticulado* AQ0FPP |

# Named Entity Recognition and Classification

- Setting
- Datasets
- Systems

# Named Entity Recognition and Classification (NERC)
# **Setting**

- NER is a subtask of Information Extraction.
- Named entities are phrases that contain the names of persons, organizations, locations, times and quantities.

[ORG U.N. ] official [PER Ekeus ] heads for [LOC Baghdad ] .

- Evaluation campaings
  - Message Understanding Conference in 1995 (MUC6)
  - Message Understanding Conference in 1997 (MUC7)
  - CONLL 2002 shared task
  - CONLL 2003 shared task

# NER example

- **NERC**

Nothing special really. Comfortable and clean but very boring decor in comparison to other **NH hotels**. I stayed in **NH** in **Brussels** and **Zurich** and I really liked them because of their modern and stylish design and big rooms. This one was just like any other hotel. Basic rooms with basic and dull decor - bit disappointing. The customer service was average. The rate was very expensive and I still had to pay for Internet and **20 euros** for breakfast!!! It was good but way overpriced! The best thing about the hotel was the location - city centre, 2min from a metro stop.

# NER example

- **Co-reference**

Nothing special really. Comfortable and clean but very boring decor in comparison to **other NH hotels**. I stayed in **NH** in **Brussels** and **Zurich** and I really liked **them** because of **their** modern and stylish design and big rooms. **This one** was just like any **other** hotel. Basic rooms with basic and dull decor - bit disappointing. The customer service was average. The rate was very expensive and I still had to pay for Internet and **20 euros** for breakfast!!! It was good but way overpriced! The best thing about **the hotel** was **the location** - city centre, 2min from a metro stop.

# NER example

- **Wikification (Named Entity Linking)**

Nothing special really. Comfortable and clean but very boring decor in comparison to **other NH hotels**. I stayed in **NH** in **Brussels** and **Zurich** and I really liked **them** because of **their** modern and stylish design and big rooms. **This one** was just like any **other** hotel. Basic rooms with basic and dull decor - bit disappointing. The customer service was average. The rate was very expensive and I still had to pay for Internet and **20 euros** for breakfast!!! It was good but way overpriced! The best thing about **the hotel** was **the location** - city centre, 2min from a metro stop.

http://en.wikipedia.org/wiki/NH_Hoteles
http://es.wikipedia.org/wiki/NH_Hoteles ... http://dbpedia.org/page/NH_Hoteles
http://en.wikipedia.org/wiki/Brussels
http://en.wikipedia.org/wiki/Zurich
http://en.wikipedia.org/wiki/Euro

# Another NER example

- **Domain extension tools**

I looked for not very expensive hotels in **Luxembourg** capital, and based on internet-info, **hotel-restaurant "Italia"** seemed to be a good choice. And **it** has appeared to meet **my** expectations. Of course, **those** that are looking for luxurious accommodation or are spoilt with everything excellent, should not stay there.

http://dbpedia.org/page/Luxembourg
http://dbpedia.org/page/Hotel-Restaurant-Italia-in-Luxembourg (**NEW**!)

- Using Named Entity Repository ...

# Named Entity Recognition and Classification

- NERC Datasets
  - CONLL 2002 datasets
  - CONLL 2003 datasets
  - BBN Corpus
  - Wikigold and WikiNER
  - German Europarl
  - JRC Names
  - Ontonotes 4.0
  - Ancora
  - Synthema Entity Knowledge Base
  - Italian Content Annotation Bank (I-CAB)
  - EVALITA 2011 NER dataset
  - SWiiT: Semantic WIkipedia for Italian
  - ...

# Named Entity Recognition and Classification

- NERC Systems
  - OpenCalais
  - BBN Identifinder
  - LingPipe
  - Stanford CoreNLP
  - Freeling
  - Illinois Named Entity Tagger
  - SuperSense Tagger
  - OpenNLP
  - C&C tools
  - GATE
  - IXA-pipeline
  - ...

# Named Entity Recognition and Classification

- Named Entity Datasets & Repositories
  - WePS (Web People Search Corpus) Datasets
  - CSWA
  - KBP at TAC
  - Cucerzan 2007
  - Fader 2009
  - Dredze 2010
  - ACEtoWiki
  - AIDA CoNLL Yago
  - TAGME Datasets
  - Illinois Wikifier Datasets
  - Wikipedia Miner
  - Google Wikipedia Concepts Dictionary
  - DBpedia
  - Freebase
  - YAGO2
  - GeoNames
  - LinkedGeoData
  - ...

# Named Entity Recognition and Classification

- Named Entity Linking Systems
  - OKKAM
  - The Wiki Machine
  - Zemanta
  - AlchemyAPI
  - CiceroLite from LCC
  - Illinois Wikifier
  - DBpedia Spotlight
  - WikiMiner
  - TAGME
  - ...

# Parsing (Syntactic Analysis)

- Setting
- PARSEVAL evaluation exercices
  - http://nlp.stanford.edu/software/stanford-dependencies.shtml
- Systems
  - RASP (John Carroll & Ted Briscoe)
  - Minipar (Dekang Lin)
  - VISL (Eckhard Bick)
  - Stanford CoreNLP
  - Freeling
  - IXA-pipeline
  - ...

# Parsing (Syntactic Analysis)

- Syntax and grammar
- Phrase structure
  - Word order
  - Syntagma, phrase, constituent
    - NP, VP, AP, head, relative clause
- Grammars
  - Syntax vs. lexicon
  - Coverage: complete, partial ...
  - Chunking, clausing, ...
  - Context-free grammars
    - Terminals, no terminals, parse trees, recursivity
    - Non-local dependencies

      *The woman who found the wallet were given a reward*

# Word Sense Disambiguation

- Setting
- WSD Tutorial (Navigli 09)
- WSD Book (Agirre & Edmonds 07)
- SENSEVAL 1, 2, 3, SEMEVAL2007, 2010, ...
- Systems
  - Knowledge-based WSD
    - Conceptual Distance (Ted Pedersen)
    - SSI (Roberto Navigli), SSI-Dijkstra (Cuadros & Rigau)
    - UKB (Soroa & Agirre)
  - Corpus-based WSD
    - GAMBL (Walter Daelemans)
    - SenseLearner (Raha Mihalcea)
    - Base Concept (Rubén Izquierdo)

# Word Sense Disambiguation
# **Setting**

- WSD is the problem of assigning the appropriate meaning (sense) to a given word in a text
- “WSD is perhaps the great open problem at the lexical level of NLP” (Resnik & Yarowsky 97)
- WSD resolution would allow:
  - acquisition of knowledge: SCF, Selectional Preferences, Predicate Models, etc.
  - improve existing Parsing, IR, IE
  - Machine Translation
  - Natural Language Understanding
  - …

# Word Sense Disambiguation
## **Setting**

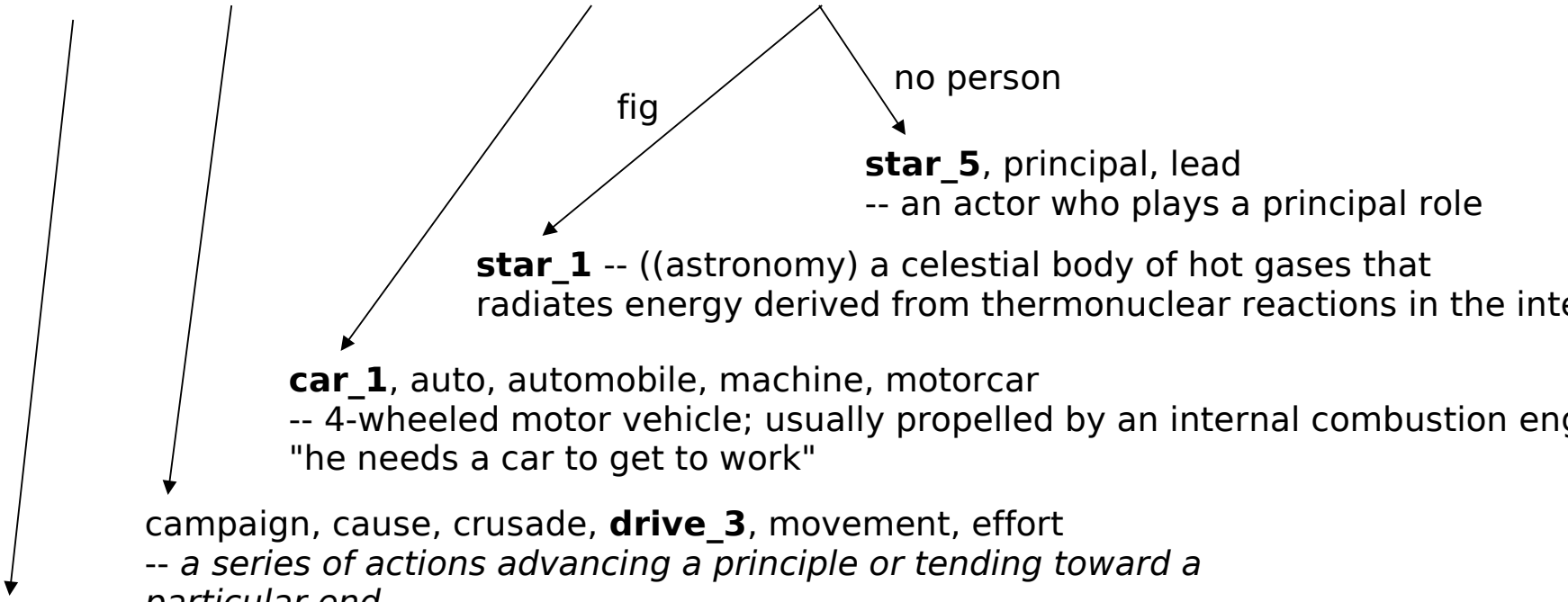
- From Financial Times

GM’s drive to make Saturn a star again

# Word Sense Disambiguation
# **Setting**

- From Financial Times

GM’s drive to make Saturn a star again



# Word Sense Disambiguation
## **Setting**

- Knowledge-Driven WSD
  - knowledge-based WSD
  - No Training Process (~ unsupervised)
  - Large scale lexical knowledge resources
    - WordNet, MRDs, Thesaurus, ...
  - 100% coverage
  - 55% accuracy (SensEval)
  - ...

# Word Sense Disambiguation
# **Setting**

- Corpus-Driven  WSD
  - statistical-based WSD
  - Machine-Learning WSD
  - Training Process (~ supervised)
    - learning from sense annotated corpora
    - (Ng 97) effort of 16 man/year per year per language
  - no full coverage
  - 70% accuracy (SensEval)

# Coreference Resolution

- Setting
- Datasets
- Systems

# Coreference Resolution

- **Co-reference** occurs when multiple expressions in a sentence or document refer to the same thing

- Mary said she would help me.

- I saw Scott yesterday. He was fishing by the lake.

# Coreference Resolution

- Datasets
  - MUC-6 (1995) and MUC-7 (1997)
  - ACE (2002 -)
  - Ontonotes
  - Ancora-CO
  - Corea
  - ...

# Coreference Resolution

- Systems
  - GUITAR
  - Bart
  - Illinois coreference Package
  - ARKref
  - Reconcile
  - MARS
  - CherryPicker
  - Stanford CoreNLP
  - RelaxCor
  - JavaRAP
  - IXA-pipeline
  - ...

# Semantic Role Labelling

- Setting
  - SRL Tutorial (Lluís Màrquez 05)
- Datasets
  - CONLL'04 shared task
  - CONLL'05 shared task
- Systems

# Semantic Role Labelling
# **Setting**

- SRL is the problem of recognizing and labelling semantic roles of a predicate
- A **semantic role** in language is the relationship that a syntactic constituent has with a predicate.
- Typical semantic arguments include:
  - Agent, Patient, Instrument, etc.
- and also adjunctive arguments:
  - Locative, Temporal, Manner, Cause, etc.
- Useful for answering "Who", "When", "What", "Where", "Why", etc.
  - IE, QA, Summarization and Semantic Interpretation

# Semantic Role Labeling
# **Setting**

- From PropBank

  [A0 He ] [AM-MOD would ] [AM-NEG n't ] [V **accept** ][A1 anything of value ] from [A2 those he was writing about ] .

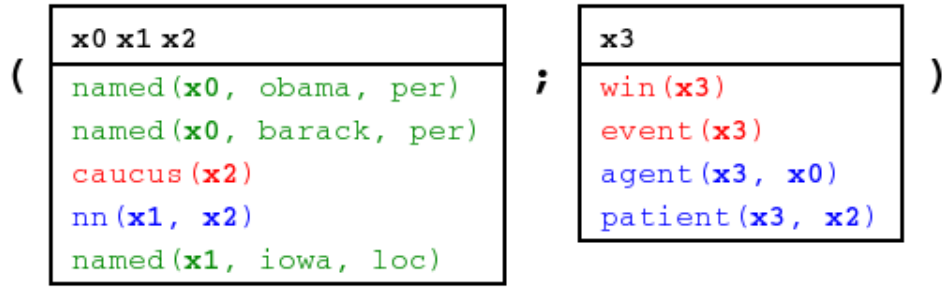
- Roleset
  - V: verb
  - A0: acceptor
  - A1: thing accepted
  - A2: accepted-from
  - A3: attribute
  - AM-MOD: modal
  - AM-NEG: negation

# Semantic Role Labelling

- Systems
  - Using PropBank rolesets ...
    - **Assert** http://cemantix.org/software/assert.html
    - **Illinois** Semantic Role Labeler
    - **SwiRL** http://www.surdeanu.name/mihai/swirl/index.php
    - **Senna** http://ml.nec-labs.com/senna
    - **MATE** tools ... http://barbar.cs.lth.se:8081
  - Using FrameNet rolesets ...
    - **Shalmanesser** ...
      - http://www.coli.uni-saarland.de/projects/salsa/shal
    - **SEMAFOR** http://www.ark.cs.cmu.edu/SEMAFOR
    - **LTH** ... http://nlp.cs.lth.se/software/semantic_parsing_framenet_frames
    - ...

# NLU

- Towards NLU
  - Boxer: ... http://svn.ask.it.usyd.edu.au/trac/candc/wiki/boxer ...
  - ...



# NLP suites

- Other
  - MALLET ... http://mallet.cs.umass.edu/ ...
  - Infomap ... http://infomap-nlp.sourceforge.net/ ...
  - DIVISI ... http://csc.media.mit.edu/divisi ...

  - BootCat ... http://bootcat.sslmit.unibo.it/ ...
    - SketchEngine ... http://sketchengine.co.uk ...
  - Webcorp ... http://www.webcorp.org.uk ...
  - ...

# NLP suites

- Complete suites for NLP
  - GATE ... http://gate.ac.uk ...
  - NLTK ... http://www.nltk.org/ ...
  - LingPipe ... http://alias-i.com/lingpipe/ ...
  - C&C tools ... http://svn.ask.it.usyd.edu.au/trac/candc/wiki
  - Freeling ... http://nlp.lsi.upc.edu/freeling/ ...
  - Stanford CoreNLP ... http://nlp.stanford.edu/software/corenlp.shtml
  - IXA-pipeline ... https://github.com/ixa-ehu
  - ...

# Basic NLP Tools

German Rigau i Claramunt
german.rigau@ehu.es
IXA group
Departamento de Lenguajes y Sistemas Informáticos
UPV/EHU