PAPEL:
a dictionary-based lexical ontology for Portuguese

Hugo Gonçalo Oliveira, Paulo Gomes, Nuno Seco
Linguateca, node of Coimbra, DEI-FCTUC, CISUC

Diana Santos
Linguateca, node of Oslo, SINTEF ICT

Universidade de Coimbra
Faculdade de Ciências e Tecnologia
Departamento de Engenharia Informática

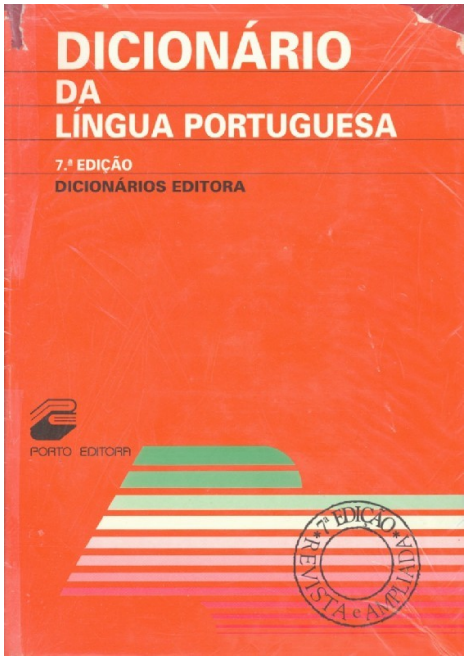
Knowledge and Intelligent Systems Laboratory
Cognitive and Media Systems Group
Centre of Informatics and Systems of the University of Coimbra

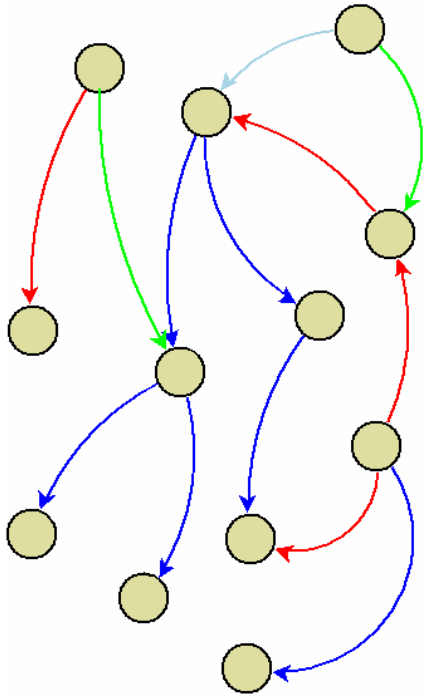
Contents

- Project objectives
- Related work
- Building PAPEL
- Further work/possible directions
- Availability

Project objectives

- PAPEL stands for Palavras Associadas Porto Editora Linguateca





MRDs as a source of semantic knowledge

- Since 1970's (Calzolari 1977, Amsler 1981, …)
- Restricted and predictable vocabulary
- Typical definition structure: genus + differentia
  - *genus:* the superordinate concept
  - *differentia:* properties for distinction between instances of the same superordinate concept.

Similar resources

- Princeton WordNet (Miller 1990)
- Microsoft's MindNet (Richardson et al. 1998)
- FrameNet (Baker et al. 1998)
- WordNet.BR (Dias da Silva 2004)
- WordNet.PT (Marrafa et al. 2006)

Building PAPEL

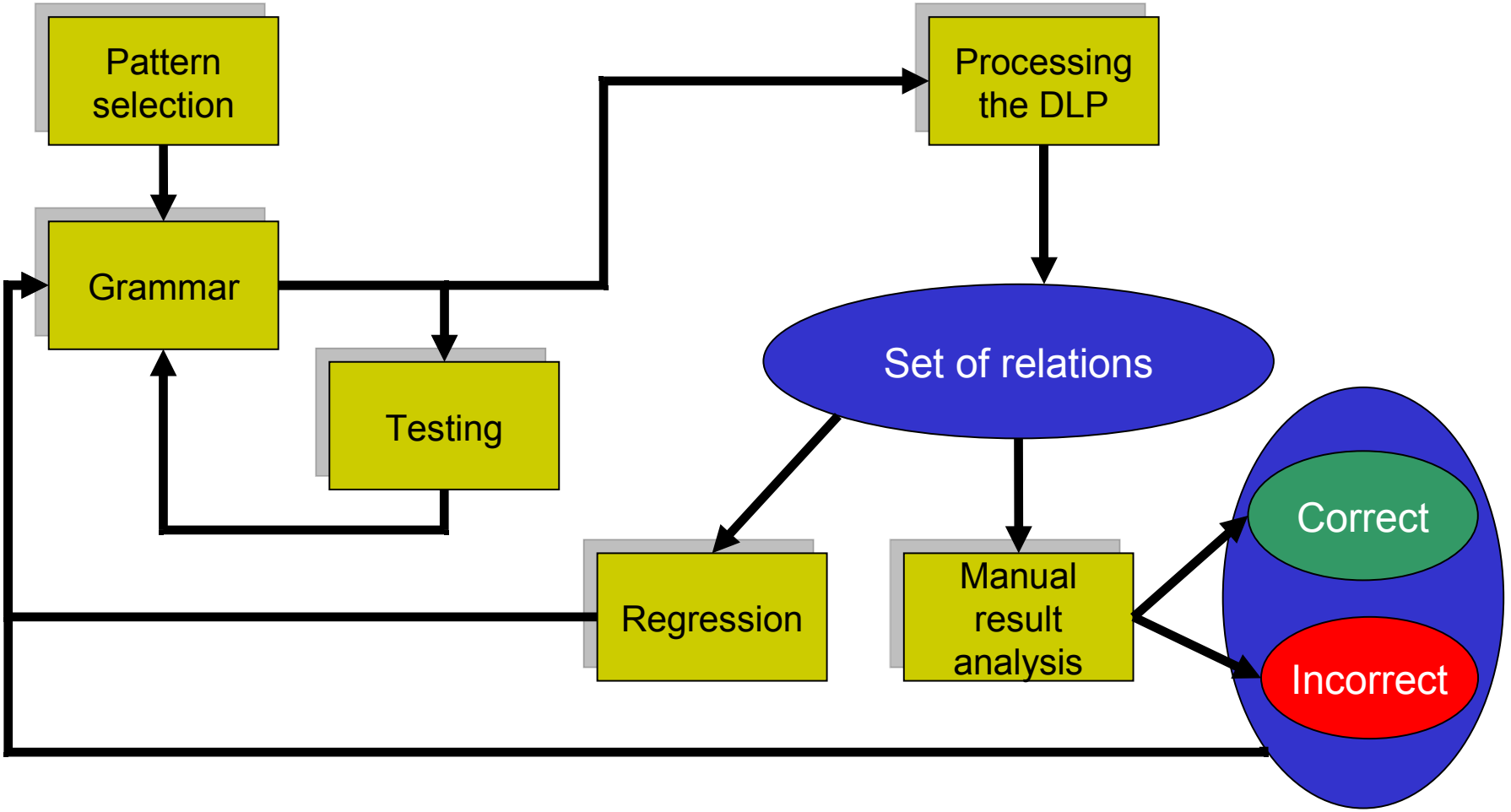
Examples of relations to include:

Relation	Read	Example
HIPERONIMO-DE(X, Y)	X is a HIPERONIMO-DE Y	HIPERONIMO-DE(animal, cão)
HIPONIMO-DE(X, Y)	X is a HIPONIMO-DE Y	HIPONIMO(cão, animal)
CAUSADOR-DE(X, Y)	X is a CAUSADOR-DE Y	CAUSADOR-DE(vírus,doença)
RESULTADO-DE(X, Y)	X is a RESULTADO-DE Y	RESULTADO-DE(doença, vírus)
MEIO-PARA(X, Y)	X is a MEIO-PARA Y	MEIO-PARA(chave, abrir)
FINALIDADE-DE(X, Y)	X is a FINALIDADE-DE Y	FINALIDADE-DE(abrir, chave)
PARTE-DE(X, Y)	X is a PARTE-DE Y	PARTE-DE(roda, carro)
INCLUI(X, Y)	X INCLUI Y	INCLUI(carro, roda)
LOCAL-DE(X, Y)	X is a LOCAL-DE Y	LOCAL-DE(restaurante, comer)
OCORRE-EM(X, Y)	X OCORRE-EM Y	LOCALIZADO-EM(comer, restaurante)

Building PAPEL

- Quantitative studies about the patterns used in the definitions.
- Using the chart parser PEN with specific grammars to extract relations [http://www.linguateca.pt/Coimbra]
- Extraction of relations between entities, from definitions:
  - `Y1 - tipo de X1` → HIPERONIMO-DE(X1, Y1)
  - `Y2 - formado por um X2` → PARTE-DE(X2, Y2)

Building PAPEL



Detailed example

- Causation relation:
  - relation between an agent (the causer) and a result (the caused)
  - CAUSADOR-DE and its inverse RESULTADO-DE.
  - Based on verbs: *causar, originar, provocar, produzir, gerar, motivar, suscitar.*

Detailed example

- Simplified rules for CAUSADOR-DE:
  - {causad|originad|provocad|produzid|gerad|motivad|suscitad} {o|os|a|as} FREQ* PREP CAUSADOR
    - `concussão (s.f.) - choque violento` **`originado por`** `uma explosão`
      - **CAUSADOR-DE**(explosão, concussão)
  - devido {a|ao|à|às|aos} CAUSADOR
    - `engasgo (s.m.) - incapacidade de respirar` **`devido a`** `obstrução da garganta`
      - **CAUSADOR-DE**(obstrução , engasgo)

Detailed example

- Simplified rules for RESULTADO-DE:
  - que {causa | origina | provoca | produz | motiva | gera | suscita} RESULTADO
    - `incomodidade (s.f.) - o` **`que causa`** `mal-estar ou desconforto`
      - **RESULTADO-DE**(mal-estar, incomodidade)
      - **RESULTADO-DE**(desconforto, incomodidade)
  - {causar | originar | provocar | produzir |motivar | gerar | suscitar} RESULTADO
    - `germe (s. m.) - microrganismo susceptível de` **`provocar`** `uma doença`
      - **RESULTADO-DE**(doença, germe)

Detailed example

- Examples of current problems:
  - Specific relations inside the definition
    - `estetoscópio (s.m.) – instrumento para auscultar a respiração, as batidas do coração e outros` **`sons produzidos pelo corpo`**`.`
      - CAUSADOR-DE(corpo, estetoscópio)
  - Identifying the result
    - `erinose (s.f.) - doença das videiras originada por um ácaro que ataca as folhas destas plantas,` **`provocando`** `a formação de umas galhas felpudas`
      - RESULTADO-DE(formação, erinose)?
      - RESULTADO-DE(galhas, erinose)?

Evaluation of results

Precision

- Manual relation evaluation
- Online interactive game, like RealMics [http://www.realmics.com]

Relation name	Runs	Hits	Tagged	Tagged correct	Tagged incorrect	Precision
CAUSADOR-DE	6	5955	3551	3440	111	0.97
RESULTADO-DE	6	1870	1190	1101	89	0.93
FINALIDADE-DE	2	1432	284	275	9	0.97
MEIO-PARA	2	813	326	302	24	0.93
INCLUI	3	585	322	315	7	0.98
PARTE-DE	2	811	202	171	31	0.85

Number of definitions in the whole DLP = 237,246

Evaluation of results

- Recall (always hard to measure!)
  - randomly choose a subset of the dictionary entries and create the results for them manually (too much work, not done so far…)
  - use independent lists by gathering / eliciting from people relations between lexemes. If the information in the lists is enconded in the dictionary, we can use it as a golden standard.
  - publicly available wordnets for other languages + machine translation + some manual cleaning (prone to a lot of problems)

Further work

- Improvements to PEN + grammars
- Weigh the use of external resources (morphological analyser, broad-coverage parser…)
- Study the presence of modifiers (*grande*, *forte*, *ligeira*, …) in the definitions
- Word sense ambiguation (Dolan 1994)
- Evaluation

Further work

- Learn from the ReRelEM (1) experience
  - Network representation
  - Expansion of the network
    - obtaining inverse relations
      - PART-OF(A, B) → CONTAINS(B, A)
    - applying transitivity rules:
      - PART-OF(A,B) + PART-OF (B,C) → PART-OF (A,C)
  - Use of the golden collection for pattern mining?

1 – Evaluation of semantic relations among named entities (HAREM track)

Availability

- Reports and presentations are available in *http://www.linguateca.pt/PAPEL*
- First version: publicly available in June 2009 (according to contract with Porto Editora)
- Further ontology creation projects in the scope of future projects or PhD theses.

The end...

Acknowledgement

- *This work was done in the scope of Linguateca, contract nº339/1.3/C/NAC, project jointly funded by the Portuguese Government and the European Union.*

POS_Conhecimento
Programa Operacional Sociedade do Conhecimento

União Europeia
FEDER

UMIC
Agência para a Sociedade do Conhecimento

FCCN
Fundação para a Computação Científica Nacional

Universidade de Coimbra
Faculdade de Ciências e Tecnologia
Departamento de Engenharia Informática

Knowledge and Intelligent Systems Laboratory
Cognitive and Media Systems Group
Centre of Informatics and Systems of the University of Coimbra