Para quê gastar 3 dias em trabalho manual quando podemos fazer (em 3 meses) uma ferramenta que o automatize?

ou

Lutando por um NLP operating system

J.João Dias de Almeida (jj@di.uminho.pt)

11 de Setembro de 2008

Linguateca

Missão

Melhorar a qualidade do PLN do Português, através de atenuar as dificuldade com que se deparam que os investigadoes e desenvolvedores desta área.

- Disponibilizando recursos e ferramentas que permitam processamento sofisticado do Português.
- Catalogando e monitorizando a area
- Organizando actividades de avaliação (conjunta)

Missão

... de acordo com as nossa preocupações e crenças:

- Language processing techniques:
  - Domain specific languages (DSL)
  - grammar based
  - rule based
- Formal approach applied to informal, noisy domains
- Scripting and fast prototype
- Deal with real size resources

Inserido na realidade académica (U.Minho / DI)

e integrando coisas divertidas

PLN operating system

```
com1 < a > b                          b = f(a)
com1 < a | com2 | com3 > b            b = com3(com2(com1(a)

estruturas mais complexas             grafo de cálculo -- Makef
```

```
texto = seq(linha)          -- texto = frases, palavras, lemas
grep                        -- nlgrep
yacc                        -- nlyacc
```

Algumas ferramentas Natura-Linguateca

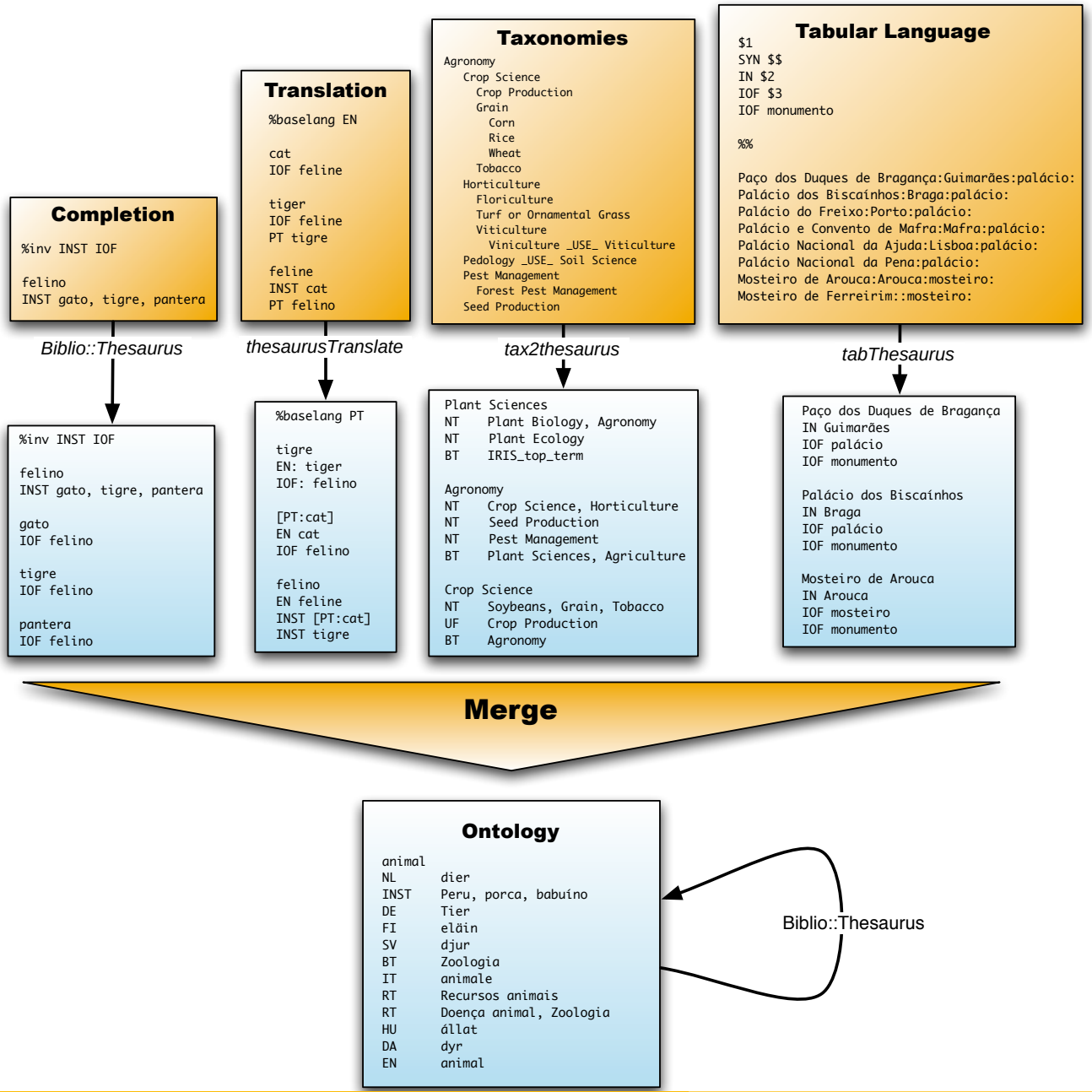
- **TerminUM**: Terminology from bilingual resources:
  - NATools: set of tools for parallel corpora:
    - aligners
    - probabilistic translation dictionary extractor
    - parallel corpora client-server
    - bilingual terminology extractor
  - Parguess: find bibtexts in the WEB

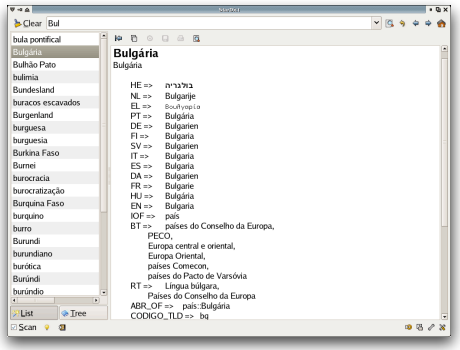
Algumas ferramentas Natura-Linguateca (2)

- Recursos e ferramentas de língua:
  - JSpell: analisador morfológico
  - Chuveiro de Dicionários: um sistema para geração de dicionários de correctores ortográficos Firefox, Thunderbird, OpenOfice, aspell, etc para PT
  - DPL: Dictionary programming language
  - DAC
  - Lingua::StarDict::Gen: geração de dicionários StarDict

Algumas ferramentas Natura-Linguateca (3)

- **T_2O**: Ontology algebra:
  - Biblio::Thesaurus
  - TabularThesaurus
- General Support tools:
  - Makefile::Parallel: DSL for large scale computing using cluster
  - Text::RewriteRules: DSL for Textual rewriting systems in Perl context
  - XML::TMX: processing Translation Memories
  - XML::DT: processing HTML and XML files


Completion
%inv INST IOF
felino
INST gato, tigre, pantera
Biblio::Thesaurus
%inv INST IOF
felino
INST gato, tigre, pantera
gato
IOF felino
tigre
IOF felino
pantera
IOF felino
Translation
%baselang EN
cat
IOF feline
tiger
IOF feline
PT tigre
feline
INST cat
PT felino
thesaurusTranslate
%baselang PT
tigre
EN: tiger
IOF: felino
[PT:cat]
EN cat
IOF felino
felino
EN feline
INST [PT:cat]
INST tigre
Taxonomies
Agronomy
Crop Science
Crop Production
Grain
Corn
Rice
Wheat
Tobacco
Horticulture
Floriculture
Turf or Ornamental Grass
Viticulture
Viniculture _USE_ Viticulture
Pedology _USE_ Soil Science
Pest Management
Forest Pest Management
Seed Production
tax2thesaurus
Plant Sciences
NT Plant Biology, Agronomy
NT Plant Ecology
BT IRIS_top_term
Agronomy
NT Crop Science, Horticulture
NT Seed Production
NT Pest Management
BT Plant Sciences, Agriculture
Crop Science
NT Soybeans, Grain, Tobacco
UF Crop Production
BT Agronomy
Tabular Language
$1
SYN $$
IN $2
IOF $3
IOF monumento
%%
Paço dos Duques de Bragança:Guimarães:palácio:
Palácio dos Biscaínhos:Braga:palácio:
Palácio do Freixo:Porto:palácio:
Palácio e Convento de Mafra:Mafra:palácio:
Palácio Nacional da Ajuda:Lisboa:palácio:
Palácio Nacional da Pena:palácio:
Mosteiro de Arouca:Arouca:mosteiro:
Mosteiro de Ferreirim::mosteiro:
tabThesaurus
Paço dos Duques de Bragança
IN Guimarães
IOF palácio
IOF monumento
Palácio dos Biscaínhos
IN Braga
IOF palácio
IOF monumento
Mosteiro de Arouca
IN Arouca
IOF mosteiro
IOF monumento
Merge
Ontology
animal
NL dier
INST Peru, porca, babuíno
DE Tier
FI eläin
SV djur
BT Zoologia
IT animale
RT Recursos animais
RT Doença animal, Zoologia
HU állat
DA dyr
EN animal
Biblio::Thesaurus



OWL
Web Ontology
Language

Biblio::Thesaurus::OWL

thesaurus2StarDict

Database
SQLite

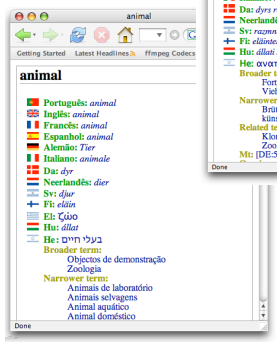
Biblio::Thesaurus::SQLite

Ontology

```
animal
NL      dier
INST    Peru, porca, babuíno
DE      Tier
FI      eläin
SV      djur
BT      Zoologia
IT      animale
RT      Recursos animais
RT      Doença animal, Zoologia
HU      állat
DA      dyr
EN      animal
```

thesaurus2tex

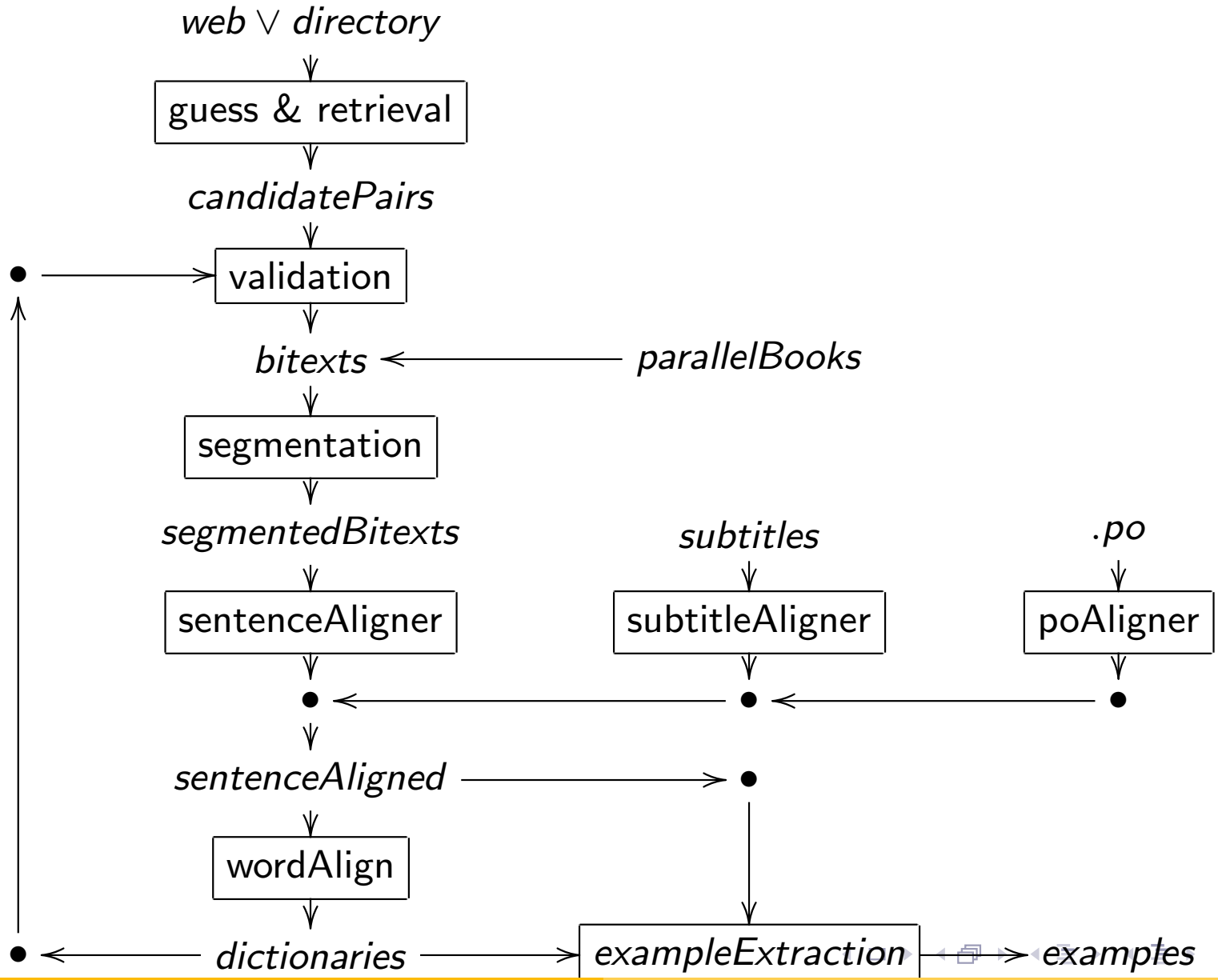
thesaurus2htmls



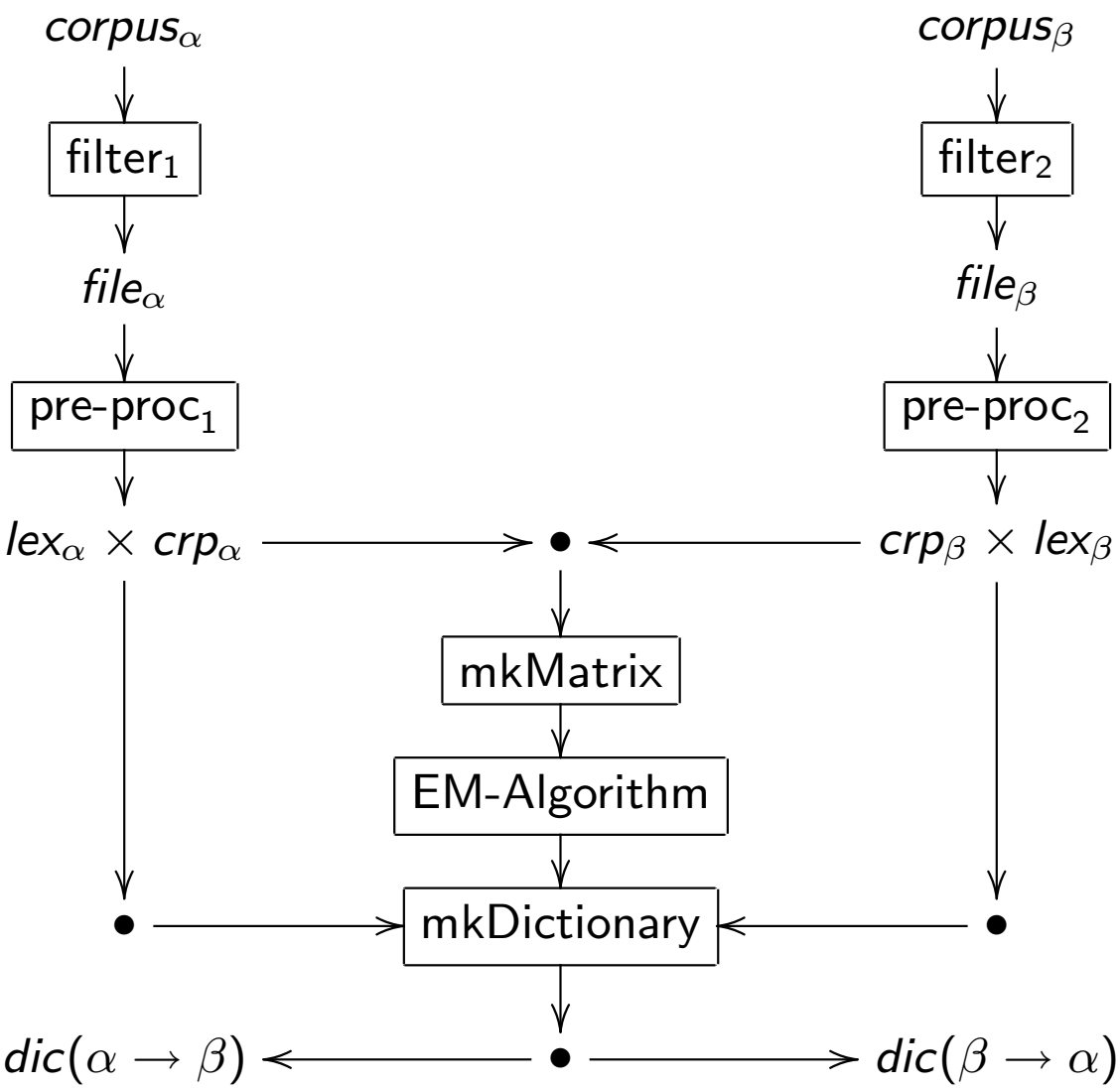
thesaurus2dict

Translation
Dictionaries

TerminUM Diagram



NATools Diagram (mais logo)



Makefile::Parallel