Port4NooJ

Portuguese Linguistic Module and Bilingual Resources for Machine Translation

Anabela Barreiro
barreiro_anabela@hotmail.com

FLUP & CLUP-Linguateca
New York University
1. Original Sources
2. Dictionaries
3. Inflectional and Derivational Description
4. Derivation
5. Syntactic-Semantic Ontology
6. Grammars
   - Morphological
   - Syntactic-Semantic
   - Disambiguation
7. Applications
   - Translation
8. Statistics
9. Future Goals
10. How to get Port4NooJ
11. Questions?
Original Sources

- OpenLogos English-Portuguese dictionary
  
  http://logos-os.dfki.de/

  OpenLogos is an open-source derivative of the Logos Machine Translation System

- NooJ linguistic environment
  
  http://www.nooj4nlp.net/

LANGUAGE RESOURCES FOR TEACHING AND RESEARCH
Anabela Barreiro
Porto, 23 April 2008
Dictionaries

Syntactic-Semantic Attributes

Inflectional Paradigm

Part of Speech

Lemma

English Transfer

---

General dictionary sample representing all PoS, variable and invariable forms

Sample of compounds in the general dictionary

Sample of the dictionary of Multiword Expressions

Sample of the dictionary of Biomedical Terms

Sample of the dictionary of Proper Names

---

Language Resources for Teaching and Research

Anabela Barreiro

Porto, 23 April 2008
Inflectional and Derivational Description

Noun Inflectional Paradigm

Verb Inflectional Paradigm

Adjective Inflectional Paradigm

Determiner Inflectional Paradigm

Adverb Inflectional Paradigm

Pronoun Inflectional Paradigm

Interrogative Pronoun Inflectional Paradigm

Nominalization Derivational Paradigm

LANGUAGE RESOURCES FOR TEACHING AND RESEARCH
Anabela Barreiro
Porto, 23 April 2008
Derivation

Verb entries:

- Identification of derivational paradigms for nominalizations (annotation \textit{DRV})

\texttt{adaptar,V+FLX=FALAR+Aux=1+INOP57+Subset=132+EN=adapt+DRV=DRV00:CANÇÃO+VSUP=fazer}

- Link to the derived noun’s support verbs (annotation \textit{VSUP})

\texttt{adaptar,V+FLX=FALAR+Aux=1+INOP57+Subset=132+EN=adapt+DRV=DRV00:CANÇÃO+VSUP=fazer}

Other Predicate Nouns need to be lexicalized:

- Identification of other predicate nouns (non-nominalizations) (annotation \textit{Npred})

\texttt{ruído,N+FLX=MENINO+Npred+ME+abs+noise+EN=noise+VSUP=fazer}
Derivation

Portuguese (Portugal)/Portuguese (Portugal) syntactic grammar consists of 2 graphs.

Eles  deram  abrigo
Syntactic-Semantic Ontology

- Representation abstract language
- Hierarchical taxonomy (sets, supersets and (sometimes) subsets)
- Based on Logos SAL ontology
- Integrated in the dictionary
- It represents both meaning (semantics), and structure (syntax)
- Over 1,000 categories
# Syntactic-Semantic Ontology

<table>
<thead>
<tr>
<th>Category</th>
<th>Mnemonic</th>
<th>Examples in English</th>
<th>Examples in Portuguese</th>
</tr>
</thead>
<tbody>
<tr>
<td>agentives</td>
<td>CO+undagt</td>
<td>See subsets</td>
<td>See subsets</td>
</tr>
<tr>
<td>software</td>
<td>CO+soft</td>
<td>routine</td>
<td>rotina, ficheiro</td>
</tr>
<tr>
<td>concrete chemical agents</td>
<td>CO+chem</td>
<td>catalyst, warhead</td>
<td>ácido sulfúrico</td>
</tr>
<tr>
<td>machines/systems</td>
<td>CO+mach</td>
<td>battery, camera</td>
<td>máquina fotográfica</td>
</tr>
<tr>
<td>vehicles</td>
<td>CO+vehic</td>
<td>truck, ship</td>
<td>automóvel</td>
</tr>
<tr>
<td>meters</td>
<td>CO+meter</td>
<td>clock, gauge</td>
<td>manómetro</td>
</tr>
<tr>
<td>communication agents</td>
<td>CO+comm</td>
<td>radio, radar</td>
<td>rádio</td>
</tr>
<tr>
<td>functionals</td>
<td>CO+undfunc</td>
<td>trinket, ornament</td>
<td>ornamento</td>
</tr>
<tr>
<td>devices/tools</td>
<td>CO+tool</td>
<td>pliers</td>
<td>alicate</td>
</tr>
<tr>
<td>fasteners</td>
<td>CO+fast</td>
<td>nail, tendon</td>
<td>prego</td>
</tr>
<tr>
<td>bearing surfaces</td>
<td>CO+surf</td>
<td>table, shelf</td>
<td>mesa</td>
</tr>
<tr>
<td>receptacles</td>
<td>CO+recp</td>
<td>bottle, barrel</td>
<td>garrafa</td>
</tr>
<tr>
<td>conduits</td>
<td>CO+cond</td>
<td>chute, artery</td>
<td>artéria</td>
</tr>
<tr>
<td>thresholds/focal points/barriers</td>
<td>CO+barr</td>
<td>wall, door</td>
<td>porta</td>
</tr>
<tr>
<td>links/bridges</td>
<td>CO+link</td>
<td>circuit, nerve</td>
<td>circuito</td>
</tr>
<tr>
<td>cloth things</td>
<td>CO+cloth</td>
<td>shirt, blanket</td>
<td>camisola</td>
</tr>
<tr>
<td>structural elements</td>
<td>CO+struc</td>
<td>spar, bone</td>
<td>osso</td>
</tr>
<tr>
<td>concretizations of verbals</td>
<td>CO+verb</td>
<td>threading</td>
<td></td>
</tr>
<tr>
<td>concretizations of mass nouns</td>
<td>CO+mass</td>
<td>acid lining</td>
<td></td>
</tr>
<tr>
<td>product/brand names</td>
<td>CO+brand</td>
<td>Windows NT</td>
<td>Windows NT</td>
</tr>
<tr>
<td>natural things</td>
<td>CO+nat</td>
<td>See subsets</td>
<td>See subsets</td>
</tr>
<tr>
<td>minute flora</td>
<td>CO+flora</td>
<td>algae, spore</td>
<td>alga</td>
</tr>
<tr>
<td>plants</td>
<td>CO+plant</td>
<td>rose, weed</td>
<td>erva</td>
</tr>
<tr>
<td>trees</td>
<td>CO+tree</td>
<td>apple, willow</td>
<td>macieira</td>
</tr>
<tr>
<td>trees/wood</td>
<td>CO+trwd</td>
<td>oak, maple</td>
<td>carvalho</td>
</tr>
<tr>
<td>misc. natural things</td>
<td>CO+mnat</td>
<td>pebble, iceberg</td>
<td>iceberg</td>
</tr>
<tr>
<td>edibles (non-mass)</td>
<td>CO+ednm</td>
<td>pork chop</td>
<td>costaleta</td>
</tr>
<tr>
<td>edibles/color</td>
<td>CO+edcol</td>
<td>orange, cherry</td>
<td>laranja</td>
</tr>
<tr>
<td>impulses/lights</td>
<td>Co+light</td>
<td>lamp, beam</td>
<td>lâmpada</td>
</tr>
<tr>
<td>blemishes/marks</td>
<td>CO+blem</td>
<td>scratch, freckle</td>
<td>sarda</td>
</tr>
<tr>
<td>classifiers</td>
<td>CO+class</td>
<td>element</td>
<td>elemento</td>
</tr>
<tr>
<td>amorphous</td>
<td>CO+amor</td>
<td>breeze, tide</td>
<td>brisa</td>
</tr>
<tr>
<td>atomistic</td>
<td>CO+atom</td>
<td>electron, atom</td>
<td>átomo</td>
</tr>
<tr>
<td>undifferentiated</td>
<td>CO+obj</td>
<td>trifle, curio</td>
<td></td>
</tr>
</tbody>
</table>

**Categories of CONCRETE nouns**
Grammars

- Morphological
- Syntactic-Semantic
- Disambiguation
- Translation
Morphological Grammar

Graph for contracted forms resulting from preposition *de* with demonstrative and personal pronouns

Morphological grammar for Portuguese contracted forms

LANGUAGE RESOURCES FOR TEACHING AND RESEARCH
Anabela Barreiro
Porto, 23 April 2008
Syntactic-Semantic Grammar

Graph for recognition and annotation of PERSON-type named entities

Concordance with annotation for PERSON-type named entities

LANGUAGE RESOURCES FOR TEACHING AND RESEARCH
Anabela Barreiro
Porto, 23 April 2008
1. a noun (N, masculine, singular, alphanumeric (alnum)), corresponding to the EN word a, that is, the letter of the alphabet;
2. a preposition (PREP), locative (Loc), corresponding to the EN preposition at;
3. a pronoun (PRO), personal (PERS), direct object (OD), third person, singular, feminine, corresponding to the EN pronoun her;
4. a determiner (DET), definite article (DEFART), feminine singular and whose canonical form is o, corresponding to the EN article the.
Applications

- Information Extraction
- Summarization
- Paraphrasing
- Pre-Editing of Texts
- Controlled Writing
- **Machine Translation**
Translation Grammar

Graph to translate simple sentences

Translation and bilingual paraphrasing of simple sentences

Concordance for Text PFSimpleSentences-3pl.nlt

<table>
<thead>
<tr>
<th>Text</th>
<th>Before</th>
<th>Seq.</th>
<th>After</th>
</tr>
</thead>
<tbody>
<tr>
<td>Elas visitam cidades.</td>
<td>Elas são homens honestos</td>
<td>Elas compram casas/they purchase houses</td>
<td>Elas são mulheres simpáticas.</td>
</tr>
<tr>
<td>Elas visitam cidades.</td>
<td>Elas são homens honestos</td>
<td>Elas compram casas/they purchase houses</td>
<td>Elas são mulheres simpáticas.</td>
</tr>
<tr>
<td>comida. Eles vendem sapatos.</td>
<td>Elas compram casas/they purchase houses</td>
<td>Elas sonham acordados. Eles sonham</td>
<td></td>
</tr>
<tr>
<td>comida. Eles vendem sapatos.</td>
<td>Elas compram casas/they purchase houses</td>
<td>Elas sonham acordados. Eles sonham</td>
<td></td>
</tr>
<tr>
<td>comida. Eles vendem sapatos.</td>
<td>Elas compram casas/they buy houses</td>
<td>Elas sonham acordados. Eles sonham</td>
<td></td>
</tr>
<tr>
<td>comida. Eles vendem sapatos.</td>
<td>Elas compram casas/they buy houses</td>
<td>Elas sonham acordados. Eles sonham</td>
<td></td>
</tr>
</tbody>
</table>

Query

6/29

LANGUAGE RESOURCES FOR TEACHING AND RESEARCH
Anabela Barreiro

Porto, 23 April 2008
## Statistics

<table>
<thead>
<tr>
<th>Dictionaries</th>
<th>Nouns</th>
<th>Verbs</th>
<th>Adjectives</th>
<th>Adverbs</th>
<th>Other PoS</th>
<th>Total (Lemmas)</th>
</tr>
</thead>
<tbody>
<tr>
<td>General Dictionary</td>
<td>30,000</td>
<td>11,000</td>
<td>2,800</td>
<td>4,700</td>
<td>11,500</td>
<td>60,000</td>
</tr>
<tr>
<td>Named Entity Dictionary</td>
<td>6,000</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>6,000 *</td>
</tr>
<tr>
<td>Multiword Expression Dictionary</td>
<td>20,000</td>
<td>10,000</td>
<td>5,000</td>
<td>5,000</td>
<td>-</td>
<td>40,000 *</td>
</tr>
<tr>
<td>Total</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>106,000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Grammars</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Portuguese-English SVC Grammars</td>
<td>---</td>
<td>2,000</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>2,000 *</td>
</tr>
<tr>
<td>Other Portuguese-English Multiword Expression Grammars</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>5,000</td>
</tr>
<tr>
<td>Total</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>7,000</td>
</tr>
</tbody>
</table>

* Not available as open source, but will be available in the annotation system
Future Goals

- Extend and Enhance dictionaries
- Enhance 6,000 existing Portuguese-English local grammars and convert over 40,000 for other language pairs
- Extend Bilingual Dictionary to multilingual with inclusion of French, Italian, Spanish and German targets (approx. 60,000 entries each language)
- Develop controlled language writing tools
- Develop Portuguese to English machine translation
How to Get Port4NooJ

Publicly available at:

http://www.nooj4nlp.net

http://www.linguateca.pt/Repositorio/Port4Nooj/
Questions?

Thank you for your attention!

Anabela