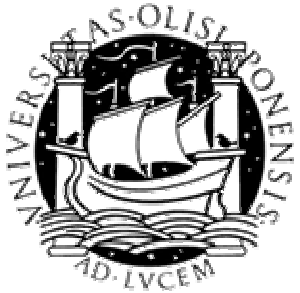
LXGram and CINTIL treebanks

Lisbon status update

António Branco

University of Lisbon



1 Lisbon Delph-in projects

- Overall effort
  - ca. 130 000 euro from April 2005
  - ca. 80 000 euro to go
- Ongoing
  - SemanticShare: grammar-based treebanking / treebank-based grammar
  - 2 years: March 2008...
  - 160 000 euro, Portuguese FCT
- Past
  - GramaXing: grammar development
  - 2 years: ...July 2007
  - 50 000 euro, Portuguese FCT

2 Management

- In tandem version development cycles
  - 3-4 months each version
  - off-synchrony by one
    - Treebank ..., **V*n***
    - Lexicon ..., V*n*, **V*n*+1**
    - Grammar: ..., V*n*, **V*n*+1**
  - ready to start v3/v4 after holidays (initial: v0)
- In-cycle progression
  - after regression check: retrain – parse – validate – adjudicate
- Last stable version
  - v2:
    - Gram&Lex: March 2009, Treebk: June 2009
  - data reported here

3 CINTIL Treebanks

- Annotation methodology
  - double annotation + adjudication
- News corpus (674 sent)
  - manual POS, lemma, infl, NER
  - repository: 11 918 sent
    - 30 tokens/sent
  - 23% parsed: 2 789
  - 6% adjudicated: 674
- Regression corpus (530 sent)
  - manual POS, lemma, infl, NER
  - repository: 555 sent
    - 7 tokens/sent aver
  - 99.8% parsed: 554
  - 95% adjudicated: 530
- 1 200 sentences adjudicated
- Inter-annotator agreement
  - Accept/reject:
    - 0.87
  - Parse tree selection:
    - 0.67 parseval tree match

4 Lexicon

- 2 step development methodology
  - lexicographic coding +
  - grammar type transposition
- Full coding approach
  - every lemma coded with all its possible syntactic profiles
- 26 000 entries / 2 000 types
  - 15 049 / 357 nouns
  - 5 054 / 1 177 verbs
  - 4 604 / 342 adjectives
  - 1 208 / 167 adverbs
  - 242 closed

- Stand alone lexicon
  - ca. 26 000 entries
  - ca. 2 000 types
  - full coverage of repository corpus
- Transposed lexicon
  - 88% entries (22 778)
  - 6% types (366)
  - 78% sentences of repository fully covered

5 LXGram grammar

- Quantitiative
  - 23 211 lines of code
    - + 6,455 comment lines
  - 3 918 types (excluding glb's)
    - 577 leaf lexical types
    - 442 lexical supertypes
  - 93 syntactic rules
  - 48 morphological rules
    - for morphology and alternances
- Shallow pre-processing
  - POS, lemmata, inflection features, NER
- News corpus coverage
  - repository: 12 000 sent, 350 000 tokens
  - 23% parsed
  - 6% correct
- PET speed
  - 70 ms/sent

- Qualitative
  - European and American variants
  - NP structure
  - Basic sentence structure
  - SVO, VSO, VOS, OVS, OSV word orders
  - preliminary clitic support
  - relatives (Subj + DO with antecedent)
  - topicalizations (Subj + DO)
  - control (Subj)
  - raising
  - verbs
    - 0-place, intransitives, transitives, ditransitives, w/ obliques, w/ sentential complements
  - nouns and adjectives w/ complements
  - completive and adverbial subordination
  - coordination
    - S, VP, NP, modifiers
  - negation
  - punctuation (via afixation)

6 Documentation

- Implementation report
  - 222 pages
  - incrementaly expanded at each new grammar version
- Comments in code
  - 6 455 lines

7 Distribution

- Downloadable
  - http://nlxgroup.di.fc.ul.pt/lxgram
- Released
  - Version March 2008
    - N.B.: the data presented here are from March 2009 version
- Next release planned
  - by the end of the project (Summer next year)

8 Team

- Mariana Avelãs
  - corpus, lexicon, annotation
- Clara Pinto
  - corpus, lexicon, annotation
- João Silva (PhD student)
  - shallow pre-processing: POS, lemma, inflection, NER
- David Raposo, and next Sérgio Castro (half-time)
  - lexical transposition, evaluation
- Francisco Costa (PhD student)
  - grammar, adjudication
- João Graça (half-time)
  - propbanking tools
- Ruben Reis (MA student), joining next Fall
- António Branco with Sara Silveira
  - coordination, workflow, versioning

9 Applications

- Temporal info processing
  - PhD Francisco Costa
- OOV's subcat
  - PhD João Silva

10 Outlook

- Short-term
  - SemanticShare project (next 12 month)
    - Corpus
      - expand with parallel corpus
    - Lexicon
      - proceed with type transposition
    - Grammar
      - enlarge coverage
- Longer-term
  - ensure continuation
  - watching for funding opportunities
  - more applications

Thank you!

Treebanking: observed agreements

- Parser supported
  - 88.53% Negra, German, Brants et al, 2000
  - 86.93% Cast3LB, Spanish, Civit et al, 2003
- Fully grammar supported
  - 96.36% Hinoki, Japanese, Bond et al, 2005

ITA

- ITA coefficient
  - S = Ao – Ae / 1 – Ae
- $S^{Rejection} = 0.83$
  - Ao = rate of parsed sentences rejected by both or accepted by both
  - Ae = binary accept/reject decision
- $S^{Parseval} = 0.67$
  - Parseval: labeled constituent similarity metric
  - Si for each sentence i
    - Ao = “crossed” F-score = $2 \times P_{(A,B)} \times P_{(B,A)} \; / \; P_{(A,B)} + P_{(B,A)}$
    - Ae = averaged “crossed” F-score over all pairs of trees from the parse forest
  - $S^{Parseval}$ obtained form Si averaged over treebank