LXGrammar and their new companion probabilistic parsers

2010 Portuguese Language Status Update

António Branco

University of Lisbon

1 Lisbon Delph-in projects

- Overall effort
  - ca. 180 000 euro since April 2005
  - ca. 30 000 euro to go, end 2010
- Ongoing: SemanticShare
  - grammar-based treebanking / treebank-based grammar
  - 160 000 euro, Portuguese FCT
  - 2 years, **extended**: March 2008 – Dec 2010
- **Pending**: EC call closed June 1
  - part for continuation of grammar&treebanking
  - 2 years: starting March 2011

2 Lines of action

- In tandem version development cycles
  - off-synchrony by one
    - Treebank ..., **V*n***
    - Lexicon ..., V*n*, **V*n*+1**
    - Grammar: ..., V*n*, **V*n*+1**
  - Initial: v0/1

- Versions
  - V2/3 reported in Barcelona Deph-in meeting 2009
    - Gram&Lex: March 2009, Treebk: June 2009
  - until then: 3-4 months each version
  - **but since then: v3/4 in development**

Lines of action ctd

- Change of timeline
  - Project end extended from March to December 2010 (same funding)
  - **Smaller than possible team but in longer time** to shorten/avoid the know-how gap to the next project
  - Currently open v3/4, started Sept 2009, will close by Dec 2010

- Change of priorities
  - New lexicon entries stoped (26 000 entries)
    - but grammar type transposition continues
  - Grammar slowed down (23 200 -> 28 600 lines of code so far)
    - error correction + support to PhD work
  - Documentation slowed down (222 -> 267 pages so far)
  - **Treebanking speeded up** (1 200 -> **3 700 adjudicated sent so far**)
    - goal: to reach 5 000 sent

Lines of action ctd ctd

- Change of tactics
  - from spiral improvement over a “closed” corpus
  - to **continuous feed with “unknown” texts** (including parallel ones)
  - BUT **question #1**: was **the grammar coverage already mature enough** to support this change of annotation tactics?

- Change of strategy
  - as the project approaches the end, seek to extract the higher short-term rewards from all materials produced so far
  - obtain the first parallel treebank PT-EN
  - induce **the first probabilistic parsers for Portuguese** (constituency + dependency)
  - BUT **question #2**: was **the treebank already large enough** to support this change of project strategy?

3 LXGram coverage

- Exploratory experiment with “unknown” texts

	Wikipedia	Público	Folha S.Paulo	Total
# sentences	66 304	30 000	30 000	126 304
average words/ sentence	25	27.5	18.6	24
# parsed sentences	20 995	8 455	11 173	40 623
% parsed	32%	28%	37%	**32%**
average parses/ parsed sentence	67	87	75	73

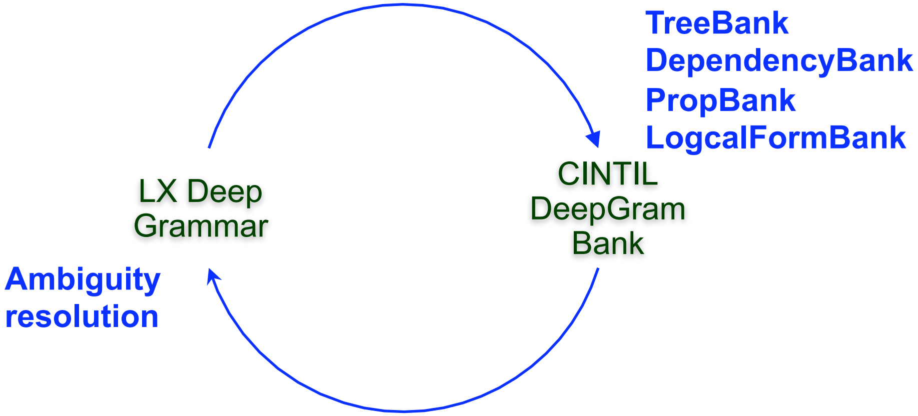
LXGram coverage ctd

- Treebanking results obtained so far

	regression test suites	**“closed” CINTIL corpus**	**“unknown ” Público**	**Total**
# sentences	875	16 000	11 900	28 775
avergae words/ sentence	7	30	27.5	28.3
# parsed sentences	874	4 338	3 618	8 830
% parsed	99.9%	27.1%	**30.4%**	30.7%
# adjudicated	787	1 757	1 130	3 674
% adjudicated	89.9%	11.0%	**9.5%**	12.8%

A virtuous circle

- **Answer #1**:
  - **yes**, LX-Grammar coverage is already large enough to support productive treebanking over open "unknown" text
  - ... with 30% parse rate (and 30% adjudication rate on top of that), it doesn't compare that bad with older sisters' parse rates:
    - 80.4% EN; 42.74% JP; 28.6% GER (Zhang, Wang, Oepen, 2009; CoNLL paper)
- Spinning a virtuous circle



4 Exogenous parsing

- Exploratory experiment with probabilistic parsing

	$f_{Parseval}$	f_{Evalb}	**POS accuracy**	**LeafAncestror**
Bikel	84.97%	73.08%	88.82%	90.48%
Stanford (Klein, Manning)	88.07%	78.75%	92.91%	91.87%
Berkeley (Petrov et al.)	**89.33**%	80.79%	91.62%	93.72%

  - 1 204 sentence treebank
  - compares very well with state of the art: 85-90% for English

- LX-Parser online: http://lxparser.di.fc.ul.pt

Exogenous parsing ctd

- Dependency parsing

	Unlabeled Attachment	Labelled attachment
KS/LR Dep (Sagae & Tsujii)	89.54%	85.01%
DeSR Dependency parser (Attardi et al.)	89.83%	85.97%
Malt parser (Nivre, Hall, Nilsson)	91.60%	87.67%
MST parser (McDonald, Lerman, Pereira)	92.19%	**90.36**%

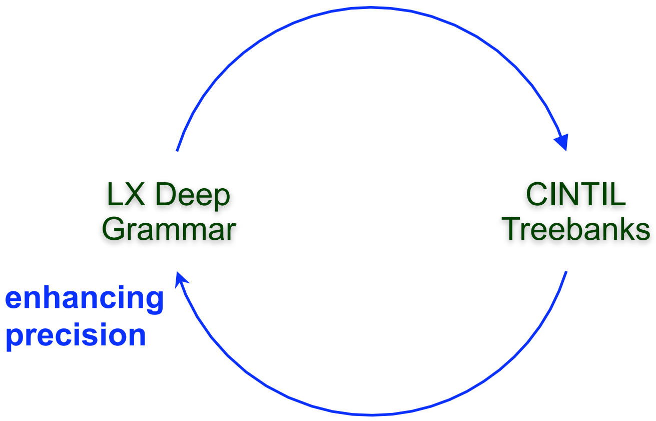
  - Compares very well with state of the art: 85-90% for English

- LX-DepParser online: http://lxdepparser.di.fc.ul.pt

Adding another virtuous circle

- **Answer #2:**
  - yes, the treebank is already large enough to support competitive probabilistic parsing

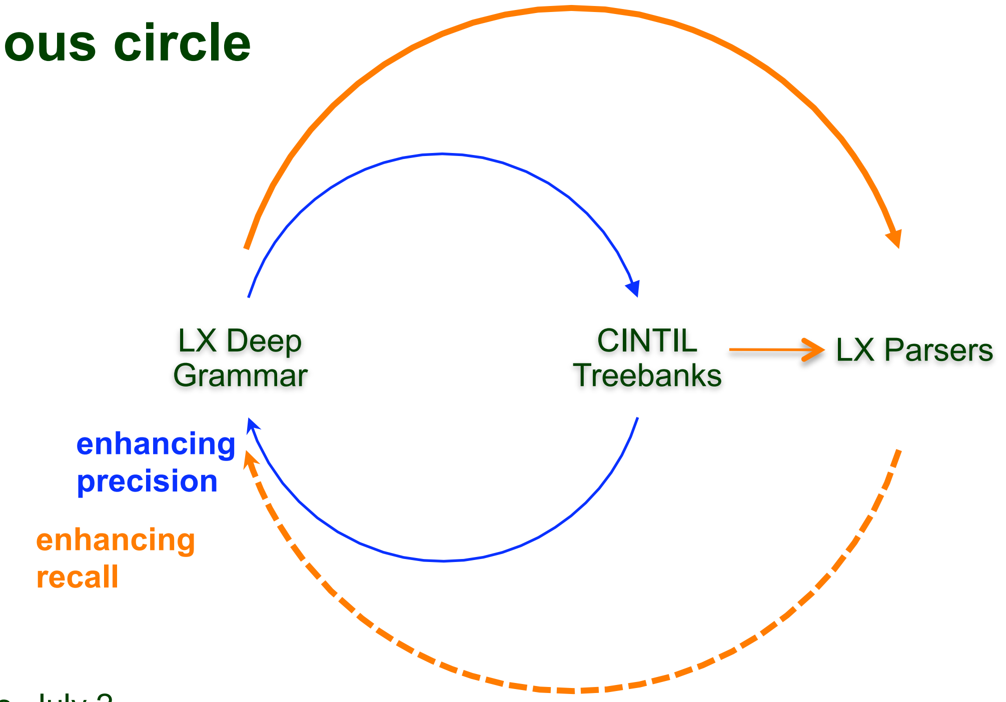
- **Adding another virtuous circle**



Adding another virtuous circle

- **Answer #2:**
  - yes, the treebank is already large enough to support competitive probabilistic parsing

- **Another virtuous circle**



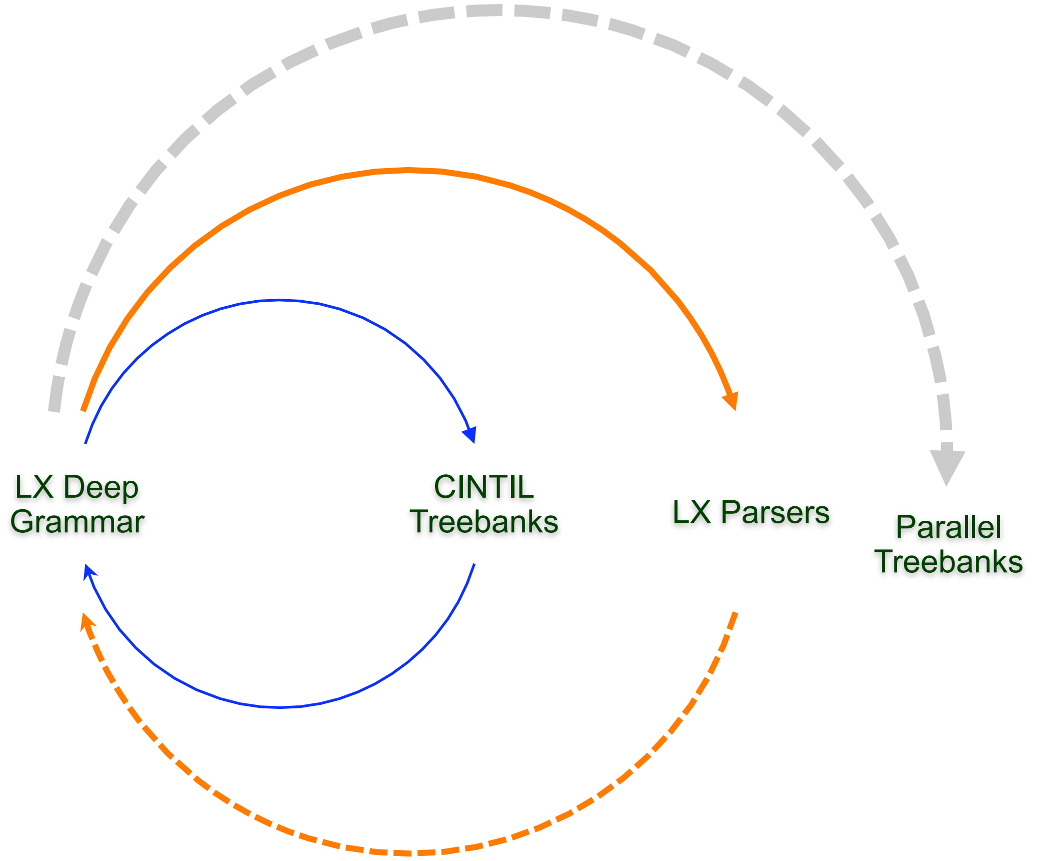
5 Parallel treebanking

- Initial exploratory experiment
  - Valia’s team announced in Barcelona meeting last year:
    - parsing WSJ with ERG
  - NLX’s team, since then:
    - translating into Portuguese + parsing with LXGram + word alignment

- First batch
  - 786 sentences from wsj02 translated
  - 272 parsed (35%); 110 adjudicated (14%)
    - **Alignment rate: around same proportion as adjudication over “open” text**
  - word alignment starting

- Second batch
  - 990 sentences from wsj00 translated

... yet another virtuous circle?



6 Distribution

- Downloadable: version March 2008
  - **Since June 2010** with **preprocessing tools also released**

- Next release planned
  - Initially planed to the end of the project (Spring 2010)
  - Rescheduled to new end of project (**end 2010**)

7 Team

- ~~Mariana Avelãs~~ **Catarina Carvalheiro**
  - corpus, lexicon, annotation
- Clara Pinto
  - corpus, lexicon, annotation
- João Silva (PhD student)
  - shallow pre-processing, constituency parsing, robustness
- ~~David Raposo~~ **Sérgio Castro** (MA student)
  - lexical transposition, evaluation, hacking around bugs in pet/lkb
- Francisco Costa (PhD student)
  - grammar, adjudication, semantics
- ~~João Graça~~ **Ruben Reis** (MA student)
  - dependency parser
- António Branco with Sara Silveira
  - coordination, workflow, versioning

8 Outlook

- Short-term
  - terminate current SemanticShare project with success (end 2010)
- Longer-term
  - ? continuation/growing with new EC project (March 2011?)
  - parallel treebanking
  - more developments/applications, more students, more advances

Thank you!