# Site Update: The University of Melbourne

Tim Baldwin, Rebecca Dridan, Andy MacKinlay, Ned Letcher

DELPH-IN Summit
25 June, 2011

# Melbourne: Overview of DELPH-IN Activities

- TreebankSearch: Web-based treebank query/exploration engine
- gDelta: visualising grammar changes
- Unsupervised parse selection
- Treeblazing and domain adaptation
- Biomedical Relation extraction with the ERG
- Supertagging

# TreebankSearch: Efficient, Scalable Treebank Querying

Web-based system for indexing/querying multilingual treebank data:

- fast/efficient
- theory/language-inspecific
- Support LPATH queries over treebank data, focusing on simple trees (not full DAGs) for now
- Support querying/dynamic rendering in a standard web browser, using standard libraries and lightweight markup

`http://hum.csse.unimelb.edu.au/ts/index`

# gDelta: Visualising the effects of grammar modifications

## [incr tsdb()] treebanking and profiling tool

- facilitates the treebanking process, allowing automatic updates
- allows comparison between two versions of a profile by
  - changes in coverage
  - changes in over-generation
  - changes in average ambiguity
  - changes in number of analyses on an item-by-item fashion
  - changes in correct analyses produced (once treebanked)

# gDelta: Visualising the effects of grammar modifications

## [incr tsdb()] treebanking and profiling tool

- facilitates the treebanking process, allowing automatic updates
- allows comparison between two versions of a profile by
  - changes in coverage
  - changes in over-generation
  - changes in average ambiguity
  - changes in number of analyses on an item-by-item fashion
  - changes in correct analyses produced (once treebanked)

*Can we automatically detect more fine-grained patterns?*

# Method

Require: the same data parsed with two versions of a grammar

1. Determine three interesting sub-sets of items:
   - parse $\rightarrow$ parse, where reading count has changed
   - parse $\rightarrow$ no parse, items that no longer parse
   - no parse $\rightarrow$ parse, items that didn't parse, and now do
2. Extract rule types as features from the analyses in each of these sets.
3. Calculate a weight for each feature and produce a feature ranking.
4. Cluster the items in each set, based on their similarity in terms of shared features.

# Output: Feature ranking

**parse → parse**

| | |
|---|---|
| 15.498 | vn-trans3-lex |
| 5.302 | rareru-obj-change-rule |
| 1.298 | simple-pass-v-morph-end-lex |
| 1.207 | hes-lex |
| 1.207 | caus-trans-obj-scope-passvmorph-end-lex |
| 1.078 | v1-v-shon-stem-lex |
| 1.052 | adv-p-lex-np-nonexh |

**parse → no parse**

| | |
|---|---|
| 76.007 | subj-zpro-ins-lrule |
| 59.877 | obj-zpro-ins-lrule |
| 35.807 | opend-obj-zpro-ins-lrule |
| 25.946 | opend-subj-zpro-ins-lrule |
| 19.419 | obj2-soc-suru-zpro-ins-lrule |
| 14.491 | obj2a-zpro-ins-lrule |
| 12.011 | obj2b-zpro-ins-lrule |

**no parse → parse**

| | |
|---|---|
| 15.498 | vn-trans3-lex |
| 1.207 | hes-lex |
| 1.078 | v1-v-shon-stem-lex |
| 0.943 | v1-c2-shon-stem-lex |
| 0.840 | hon-prefix2vn-lex |
| 0.480 | comp-prpstn-lex-questarg |
| 0.480 | wh-word-honsubj-lex |

# Output: Clusters

**parse→no parse**
Items: 71 Clusters: 4

Cluster 1: 21 items
Exemplar: 今年も残りわずかな日しかない。
(ID-9999 tanaka_ja_6)

| Feature | Weight | Cohesion | Overlap |
|---|---|---|---|
| obj-zpro-ins-lrule | 59.88 | 100% | 41% |
| sa-lexeme-infl-rule | 0.02 | 88% | 25% |
| caus-intrans-passcmorph-end-lex | 0.01 | 50% | 11% |
| vn-trans1-lex | 0.01 | 75% | 30% |
| vend-vend-rule | 0.00 | 75% | 19% |
| ⋮ | | | |