

A Suite of Linguistic Tools for Use with the Penn-II Treebank

Aoife Cahill, Mairéad McCarthy, Ruth O' Donovan,
Josef van Genabith, Andy Way

National Centre for Language Technology, School of Computing, Dublin City University,
Dublin 9, Ireland
{acahill, mccarthy, rodonovan, josef, away}@computing.dcu.ie



(Cahill et al. 2002a,b, 2003) present a methodology for compiling wide-coverage, probabilistic LFG resources (grammars, lexica and parsers) based on treebank resources which are automatically annotated with f-structure information. This poster describes the computational infrastructure to support the linguistic development of the automatic treebank f-structure annotation algorithm. This infrastructure includes tools for selective search and graphical display of treebank information, an automatic f-structure annotation tool, a constraint solver and probabilistic parsers and taggers. The tools are platform independent and can be accessed using a web-browser. (<http://www.computing.dcu.ie/research/nclt/>)

Graphical Display of Trees and Subtrees according to Rule Instances

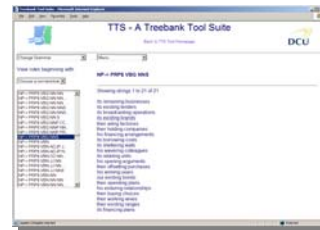


The user can select a rule (and partially specify required terminal yield) and will be presented with all subtrees in the Penn II Treebank dominated by that rule.



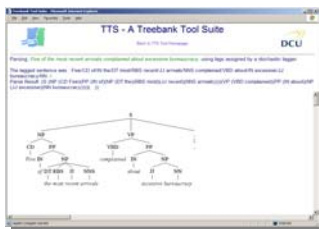
Search by rule (and yield), supports easy search for certain linguistic phenomena.

Display of the Yield of the Subtree (with and without context)



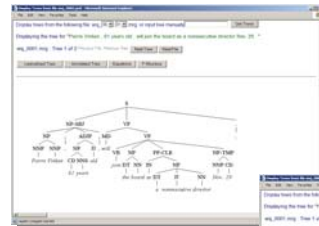
The user can view the yield of a particular Context-Free rule, either with or without context. This supports easy identification of collocations and other lexical phenomena.

Tagging and PCFG-parsing of New Input



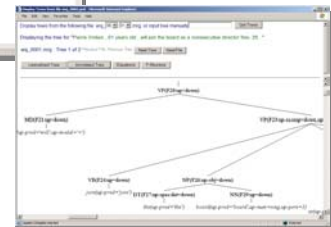
The user can input raw text which will then be tagged and parsed. The text can be tagged by either a bigram tagger, a trigram tagger, or by the parser. The parser is a probabilistic CKY chart parser, and the grammar is extracted from sections 02-21 of the Penn-II Treebank.

An Automatic Annotation Tool

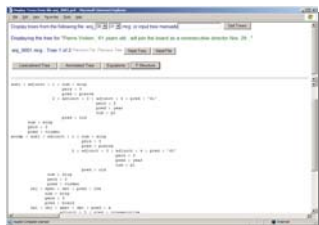


Each tree in the Penn-II Treebank can be automatically annotated with LFG f-structure information according to the algorithm described in Cahill et al. (2002a)

Each node in the tree is annotated with an LFG equation.



An Automatic Annotation Tool



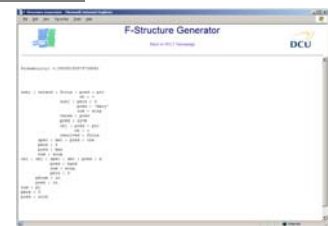
The equations on each node are collected and resolved using a constraint solver to produce an LFG f-structure. By automatically annotating each tree in the Penn-II treebank and resolving the f-structure, we semi-automatically derive a large-scale LFG resource for English.

A Wide-Coverage Probabilistic LFG Parser



The user can type in new text. It will be tagged, parsed, automatically annotated, and resolved into an LFG f-structure.

Long distance dependencies in the resulting f-structure will also be resolved. This allows us to parse raw text into proper f-structures.



References:

- Cahill A., M. McCarthy, J. van Genabith and A. Way (2002a): "Automatic Annotation of the Penn-Treebank with LFG F-Structure Information", in *Proceedings of the LREC Workshop on Linguistic Knowledge Acquisition and Representation: Bootstrapping Annotated Language Data*, Las Palmas, Spain, pp.8--15.
- Cahill, A., M. McCarthy, J. van Genabith and A. Way (2002b): "Parsing Text with a PCFG derived from Penn-II with an Automatic F-Structure Annotation Procedure", in M. Butt and T. Holloway-King (eds.) *Proceedings of the Seventh International Conference on LFG*, CSLI Publications, Stanford, CA., pp.76--95.
- Cahill A., M. Forst, M. McCarthy, R. O' Donovan, C. Rohrer, J. van Genabith and A. Way, (2003) "Treebank-Based Multilingual Unification-Grammar Development" in the *Proceedings of the Workshop on Ideas and Strategies for Multilingual Grammar Development*, at the 15th European Summer School in Logic Language and Information, Vienna, Austria, 18th - 29th August 2003