Finnish TreeBank: Grammar Definition Corpus

1. BASIC INFORMATION

1.1 Corpus composition
Dependency-annotated example sentences of the Large Grammar of Finnish

1.2 Representation of the corpus (flat files, database, markup)
One tabular CoNLL-X file

1.3 Character encoding
UTF-8

2 ADMINISTRATIVE INFORMATION

2.1 Contact person
Name: Atro Voutilainen
E-mail: atro.voutilainen@helsinki.fi

2.2 Copyright statement and information on IPR
GNU LGPL v3.0

3 TECHNICAL INFORMATION

3.1 Data structure of an entry
CoNLL-X format, token per line with partial morphological annotation and syntactic dependency-links following the running number within the sentence, the surface form and the lemma.

3.2 Corpora size (num. of tokens)
160 000 tokens, 19 000 sentences

4 CONTENT INFORMATION

4.1 Type of the corpus (monolingual/multilingual, parallel/comparable, raw/annotated)
Monolingual, annotated with partial morphology and dependency syntax

4.2 The natural language(s) of the corpus
Finnish

4.3 Domain(s)/register(s) of the corpus
Grammatical example sentences

4.4 Annotations in the corpus (if an annotated corpus)

4.4.1 Types of annotations (paragraph mark-up, sentence mark-up, lexical mark-up, syntactic mark-up, semantic mark-up, discourse mark-up)
sentences separated, lemmas, morphological tags indicate word class and inflection categories for each token, dependency-syntactic functions link dependent words to their heads

4.4.2 Tags (if POS/WSD/TIME/discourse/etc –tagged or parsed),

4.4.3 Alignment information (if the corpus contain aligned documents: level of alignment, how it was achieved)

4.5 Intended application of the corpus
Serves as a model for further grammatical analysis of Finnish

1.1. Reliability of the annotations (automatically/manually assigned) – if any
Manually annotated

5 RELEVANT REFERENCES AND OTHER INFORMATION

http://www.ling.helsinki.fi/kieliteknologia/tutkimus/treebank/index.shtml
http://www.scripta.kotus.fi/visk/etusivu.php (Iso suomen kielioppi on the web)

Atro Voutilainen, Tanja Purtonen, Satu Leisko-Järvinen, Mikaela Klami, 2010, “Suomen kielioppikorpus ja dependenssisyntaktinen kuvausmalli” [A grammar definition corpus and dependency syntactic representation of Finnish], available from the web site.