Transforming Penn Treebank Phrase Trees into (Praguian) Tectogrammatical Dependency Trees

Zdeněk Žabokrtský and Ivona Kučerová

Abstract

The aim of this article is to document a work in progress on experiments with transforming a part of the (English) Penn Treebank into tectogrammatical tree structures, similar to those which are defined in the annotation scheme of the (Czech) Prague Dependency Treebank[1]. After a brief outline of the main properties of the sentence representations used in both projects, the transformation from one representation to the other is described in detail. The cornerstones of the transformation are (i) a recursive procedure for translating the topology of a phrase tree into the Praguian tectogrammatical dependency tree topology, (ii) the procedure for functor ("thematic role") assignment, and (iii) the procedure for grammateme assignment.

By applying the transformation to the Wall Street Journal part of the Penn Treebank, the tectogrammatical tree structures for roughly 48,000 English sentences have been automatically created. Roughly 1000 trees have been manually corrected. An evaluation of the differences between the data before and after manual corrections is presented. It also allows for a general estimate of the quality of the automatically created trees.

One of the differences between tectogrammatical and phrase trees is the fact that the original sentences can be trivially reconstructed only from the latter ones. That is why we include here also a few remarks on how to generate sentences from tectogrammatical trees.

1 Introduction

1.1 Penn Treebank, Wall Street Journal

The Penn Treebank project (PTB, [6])[2] consists of about 1,500,000 tokens from English newspaper texts. The treebank bracketing style is based on constituent syntax. Not only syntactic elements, but also several types of structural reconstructions (traces) are realized on the surface. Samples of the PennTreebank bracketing and a set of frequent labels are presented in the Appendix.

The largest subpart of the PTB texts is taken from the Wall Street Journal. PTB project selected 2,499 stories from a three-year Wall Street Journal (WSJ) collection of 98,732 stories for syntactic annotation (1 million words, about 40,000 sentences). The transformation tools described below have been proceeded only on WSJ subpart of the treebank.

1.2 Prague Dependency Treebank and Tectogrammatical Tree Structures

The Prague Dependency Treebank (see [5] for references) is a research project running at the Center for Computational Linguistics[3] and the Institute of Formal and Applied Linguistics[4], Charles University, Prague. It aims at creating a complex annotation of a part of the Czech National Corpus[5]. The sentences are assigned their underlying representations in three steps

[1] The research described herein has been supported by the Ministry of Education of the Czech Republic, project LN00A063, and by the NSF Award #IIS-0121285.

[2] LDC catalog no.: LDC99T42, version 3: `http://www.ldc.upenn.edu/Catalog/LDC99T42.html`

[3] `http://ckl.mff.cuni.cz`

[4] `http://ufal.mff.cuni.cz`

[5] `http://ucnk.ff.cuni.cz`

of annotation: morphological, analytical, and tectogrammatical. The data on the first two levels, which were annotated in a semiautomatic way, consist of more than a million tokens[6]. Presently, the semiautomatic annotation on the third level has been finished for roughly 20,000 sentences.

The annotation of a sentence on the tectogrammatical level results in a *tectogrammatical tree structure* (TGTS). A TGTS is a dependency tree, whose main properties are the following: *(i)* only autosemantic (lexical, meaningful) words have a node of their own; *(ii)* the correlates of function words (i.e. synsemantic, auxiliary words) are attached as labels to the autosemantic words to which they belong (i.e. auxiliary verbs and subordinating conjunctions to the verbs, prepositions to nouns, etc.); coordinating conjunctions remain as nodes of their own; *(iii)* each node is labeled with a *functor* (arguments or theta roles, and adjuncts); *(iv)* the nodes contain a backward link to the node(s) on the second (analytical) level from which they were created, in order to retrieve the information contained there for various purposes (lemma, morphological tag, analytical function, form, surface word order etc.).

A functor represents the role of the node within the sentence, for example Actor, Patient, Addressee, Effect, Origin, various types of spatial and temporal circumstantials, Means, Manner, Condition, etc. There are roughly 60 functors. Functors provide detailed information on the relation between a node and its governing node. See Appendix B for the list of the most frequent functors.

2 Transformation Procedure

2.1 Outline

The transformation of the Penn Treebank phrase trees to the tectogrammatical trees consists of the following steps:

1. **Marking Heads** - the head is chosen in each phrase (using a program written by Jason Eisner ([2]));
2. **Lemmatisation** - a lemma is attached to each word in the sentence (using a program written by Martin Čmejrek (see Chapter 2 in [3]);
3. **Structural Transformations** - the topology of the tectogrammatical tree is derived from the topology of the PTB tree, and each node is labeled with the information from the PTB tree. In this step, the concept of head of a PTB subtree plays a key role;
4. **Functor Assignment** - a functor is assigned to each node of the tectogrammatical tree;
5. **Grammateme Assignment** - morphological (e.g. Tense, Degree of Comparison) and syntactic grammatemes (e.g. TWHEN_AFT(er)) are assigned to each node of the tectogrammatical tree. The assignment of the morphological attributes is based on Penn-Treebank tags and reflects basic morphological properties of the language. The syntactic grammatemes capture more specific information about deep syntactic structure. At the moment, there are no automatic tools for the assignment of the latter ones.

The transformation tool described in this document covers the last three steps. The tool was written in Perl and consists of roughly 1000 lines of code. The resulting tectogrammatical trees (a sample of which is available on the Internet[7]) are stored in fs-format and can be viewed using the tree editor Tred[8] ([4]).

[6] LDC catalog no.: LDC2001T10, version 1.0; http://www.ldc.upenn.edu/Catalog/LDC2001T10.html
[7] http://ckl.mff.cuni.cz/zabokrtsky/wsj2tgts/
[8] http://ckl.mff.cuni.cz/pajas/tred/

2.2 Structural Transformation

The structural transformation can be divided into two steps. First, an "initial dependency tree" (see Fig. 1) is created, in which each word and punctuation mark has its own node[9]. Second, the nodes which are not autosemantic (punctuation marks, prepositions, determiners, subordinating conjunctions, certain particles, auxiliary verbs, modal verbs) are marked as "deleted" (instead of physical deletion, they are just marked as hidden). Selected information from the deleted nodes is copied into the governing autosemantic nodes. Traces are also processed in the second step.

The topology of the initial tree is derived from the topology of the phrase tree by a recursive procedure, which has the following input arguments: phrase tree T_{phr}, initial tree T_{dep}, one particular node s_{phr} from T_{phr} – root of the phrase subtree to be processed, and node p_{dep} from T_{dep} – future parent of the tectogrammatical subtree resulting from s_{Phr} subtree. The recursion looks as follows:

1. if s_{phr} is a terminal node, then create a single tectogrammatical node n_{dep} in T_{dep} and attach it below p_{dep}; return n_{dep},

2. else (it is a nonterminal): choose the head node h_{phr} among the children of s_{phr}, run this recursive procedure with h_{phr} as the phrase subtree root argument, and it returns node r_{dep} (root of the recursively created dep. subtree); run the recursive procedure for each remaining s_{phr}'s child $n_{phr,i}$, get the subtree root $o_{dep,i}$ and attach it below r_{dep}; return r_{dep}.

Obviously, the concept of the head[10] of a phrase subtree plays the key role for the structural transformation. The notion of head used in our approach slightly differs from that of Jason Eisner's head assigning script. Therefore we occasionally use different rules for head selection (for example in case of apposition, prepositional phrases etc.).

Treating Traces. The PennTreebank annotation scheme reflects not only surface realization of a sentence, but it also contains several types of traces. Some of them can be used for generating TGTS nodes that are not realized on the surface.

- A-movement traces: marked as numbered asterisks without other letter specification (e.g. *-1); they can be transformed into several types of coreferential nodes (eg. Cor.ACT) and to nodes of so called general participants (eg. Gen.ACT); the procedure is based on the theoretical assumption that full information about predicate-argument structure is present at the tectogrammatical level; examples can be seen in the Appendix C.2.

- A'-movement traces: they are marked as numbered asterisks with letter "T" specification (e.g. *T*-1); they are used only when assigning functors to wh-word within relative clauses; in the future they could be used also for generating topic-focus articulation as one of the important pieces of information need for capturing this complex phenomenon; an example can be seen in Appendix C.1.

Other types of traces are not captured by the transformation procedure yet.

2.3 Functor Assignment

Various properties of both the phrase tree and the tectogrammatical tree are used for the functor assignment, for example:

[9] However, the initial dependency tree differs from the analytic tree as defined in the annotation scheme of the PDT. For example, the head of a prepositional phrase is not the preposition.

[10] The implementation of the head choosing part of the transformation was partly inspired by a code written by Christian Korthals for similar purposes.

- part-of-speech tags can be used in certain cases; for instance, if a word was tagged as PRP$ (possessive pronoun), then the functor APP (appurtenance) is assigned (PRP$ → APP for short; see the Appendix for the full tag sets), JJ → RSTR, JJR → CPR, etc.
- function tags: BNF → BEN, DTV → ADDR, LGS → ACT, etc.
- lemma: “not” → RHEM, “only” → RHEM, “both” → RSTR, “very” → EXT, etc.

2.4 Grammateme Assignment

Various properties of both the phrase tree and the tectogrammatical tree are used for the grammateme assignment, for example:

- in the case of nouns, number can be derived from the POS-tag (NN and NNP singular, NNS and NNPS plural)
- in the case of certain pronouns, number and gender can be derived from their lemma (“she” FEM SG, etc.)[11]
- the degree of comparison for adjectives and adverbs can be derived from their POS-tag (e.g. JJS → SUP) or from deleted function words (e.g. *more interesting* → COMP)
- tense is derived either from the POS-tag (e.g. VBZ → present) or from the combination of (deleted) auxiliary verbs
- certain grammatemes obtain automatically only their default value (e.g. IT0 for iterativeness).

3 Node Attributes

When the tectogrammatical trees are being created, each node is equipped with many attributes. Some of them are defined within the tectogrammatical level of language description (trlemma, functor, grammatemes). On the other hand, many of them have only different technical functions and do not belong to the (theoretical) tectogrammatical representation as such.

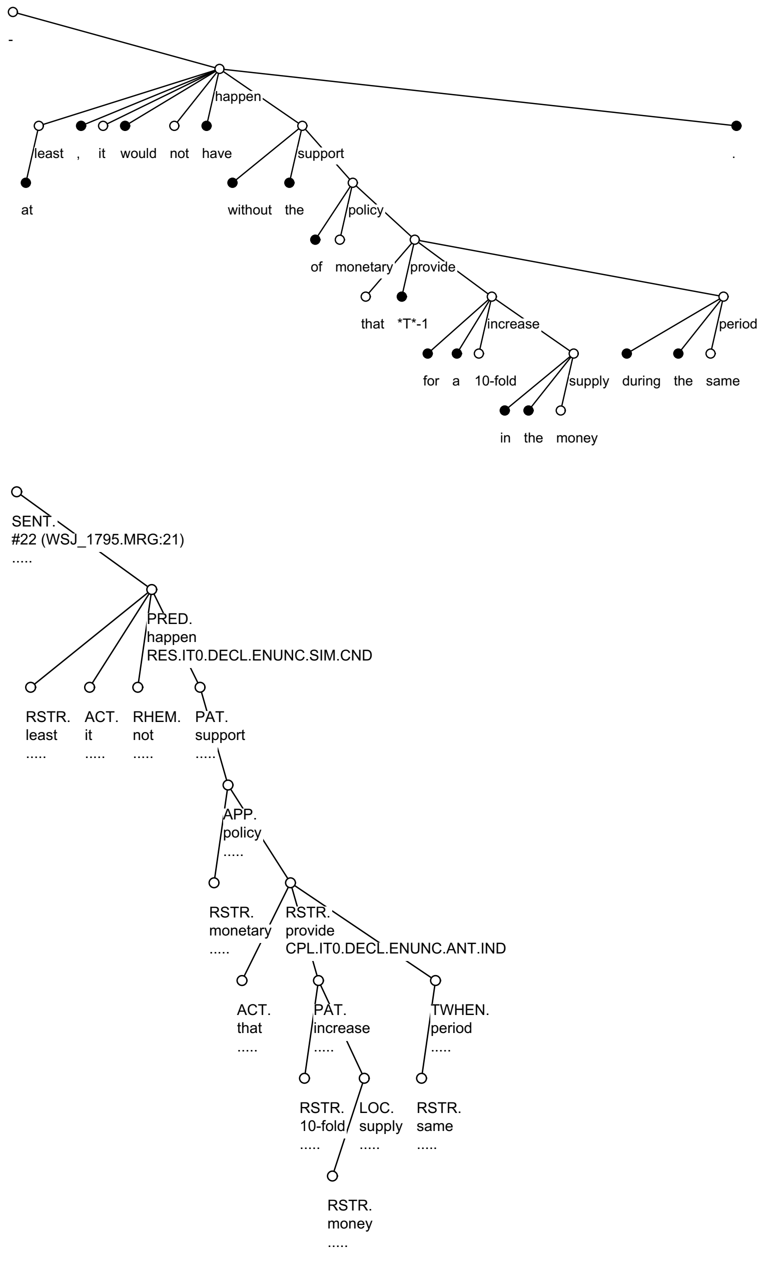
3.1 Technical Attributes

1. FORM - original word form;
2. FW (function word) - word form of a (hidden) preposition or subordinating conjunction attached below the given node;
3. X_PHRASE_SEQUENCE - sequence of labels of non-terminal nodes (of the phrase tree), which “collapsed” into one node of the tectogrammatical tree; the labels are separated by “;”; e.g.: `NN;NP~;PP-TMP`;
4. X_MODALVERB - if a hidden node with a modal verb is governed by the given node, then the word form of the modal verb is copied into this attribute of the autosemantic verb; this grammateme is used for DEONTMOD grammateme assignment;
5. X_AUXVERB_FORMS - the same thing, but with auxiliary verb forms; this attribute is used for verbal grammatemes assignment;

[11] Note that the “surface lemma” of a pronoun might differ from its tectogrammatical lemma, e.g.: “myself” → “I”, “she” → “he”, etc.

(a) **input data** – the original PTB bracketing enriched with head markers and lemmas:

```
WSJ_1795.MRG:21::(S (ADVP (@IN @at at) (JJS @least least)) (, @, ,) (NP~-SBJ (@PRP @it it))
(@VP (MD @would would) (RB @not not)(@VP~(VB @have have) (@VP~ (@VBN @happened happen) (PP (@IN
@without without) (NP~ (@NP (DT @the the) (@NN @support support)) (PP (@IN @of of) (NP~ (@NP
(JJ @monetary monetary) (@NN @policy policy)) (SBAR (@WHNP-1 (@WDT  @that that)) (S~ (NP~-SBJ
(@-NONE- @*T*-1 *T*-1)) (@VP (@VBD @provided provide) (PP-CLR (@IN @for for) (NP~ (@NP (DT @a a)
(JJ @10-fold 10-fold) (@NN @increase increase)) (PP-LOC (@IN @in in) (NP~ (DT @the the) (NN
@money money) (@NN @supply supply))))) (PP-TMP (@IN @during during) (NP~ (DT @the the) (JJ @same
same) (@NN @period period))))))))))))) (. @. .))
```



(b) **the initial dependency tree**: each word or punctuation mark has its own node; the (technical) root node is added. Nodes to be deleted (prepositions, determiners, modal and auxiliary verbs, punctuation marks, A'-movement trace) are depicted as black circles.

(c) **the resulting tectogrammatical tree**: nodes which are not autosemantic are deleted; values of functors and grammatemes are assigned (only selected verbal grammatemes are visible in the figure).

Figure 1: Process of creation of the tectogrammatical tree structure of the sentence "*At least, it would not have happened without the support of monetary policy that provided for a 10-fold increase in the money supply during the same period.*"

6. X_AUXVERB_LEMMAS - the same thing, but with auxiliary verb lemmas; this attribute is used for verbal grammatemes assignment;

7. X_DETERMINER - word form of a hidden child node with a determiner (note: “this”, “these” etc. are marked as determiners in the PTB, but we treat them as adjectives);

8. X_WSJ_ID - word identifier (format: sentence_id/word_number), e.g.: `WSJ_1795.MRG:21/13`;

9. X_TRANSLATION - translation of the lemma into Czech, which should ease the manual annotation in the case of complicated sentences (for Czech annotators, obviously).

3.2 Genuine Tectogrammatical Attributes

Note: Some non-essential distinctions between Czech and English can be found even in functor assignment, but much more serious distinctions are expected in the assignment of grammatemes, because the sets of values of grammatemes are much more language dependent. The theoretical distinctions between the tectogrammatical level for English and the tectogrammatical level for Czech have not been properly studied yet, therefore we had to use the tectogrammatical tag set as developed for Czech. This fact might be the source of certain representational inadequacies.

3.2.1 Morphological grammatemes assigned by the automatic procedure

- for verbs and deverbal forms (e.g. gerunds)
  - ASPECT
    - **PROC** - processual, i.e. analogical to the Czech imperfective form; *economist who use.*PROC *the total employment figures*
    - **CPL** - complex, i.e. analogical to perfective form; *the trade gap is expected to widen.*CPL *st.*
    - **RES** - resultative; *experts are thought to have risen.*RES *strongly in August*
  - ITERATIVENESS
    - **IT0** - *economists said.*IT0*: Exports are...*
    - **IT1** - assigned only manually

- for finite verbs only
  - SENTMOD (mode of a sentence)
    - **ENUNC** - indicative mode of the clause (applicable also for relative clauses)
    - **EXCL** - exclamatory; assigned only manually
    - **DESID** - optative; assigned only manually
    - **IMPER** - imperative; assigned only manually
  - VERBMOD (mode of a finite verb)
    - **IND** - indicative mode of the verb; *economist who use.*IND *the total employment figures*
    - **CND** - conditional form of the verb: *they could arrive.*CND *only on Monday*
    - **IMPER** - imperative; assigned only manually
  - DEONTMOD (modality of a verb)
    - **DECL** - non-modal form; *he came.*DECL *on Monday*
    - **DEB** - debitive; *he must come.*DEB

        - **HRT** - hortative; *he should come.*HRT
        - **VOL** - volitive; *he wants to come.*VOL
        - **POSS** - possibilitive; *he can come.*POSS, *he would be able to come.*POSS
        - **PERM** - permissive; *he may come.*PERM
        - **FAC** - facultative; assigned only manually
    - Tense
        - **SIM** - simultaneous; *he wants to come.*SIM; *he has not done.*SIM *it yet*
        - **ANT** - anterior; *he wanted to come.*ANT; *he had not done.*ANT *it before entering the university*
        - **POST** - posterior; *he will come.*POST *on Monday*; *he will have done.*POST *it by Monday*

- for nouns and pronouns
    - Number
        - **SG** - *he met one girl.*SG
        - **PL** - *he met girls.*PL
- only for pronouns
    - Gender
        - **ANIM** - *he*; *his*
        - **FEM** - *she*; *her*
        - **NEUT** - *it*; *its*
- for adjectives and adverbs
    - Degcmp (Degree of Comparison)
        - **POS** - positive; *small,well*
        - **COMP** - comparative; *smaller*; *more interesting*
        - **SUP** - superlative; *smallest*; *the most interesting*

3.2.2 Values of the structural grammateme memberof

- memberof
    - **CO** - at all conjoined items (CONJ) except common dependents; *all.*NIL *boys.*CO *and girls.*CO *came to the party*
    - **AP** - at items of an apposition; *John Benjamin.*AP*, 45.*AP*, was assigned …*

4 Manual Annotation

In order to gain a "gold standard" (high-quality data set), roughly 1,000 sentences have been manually corrected after the automatic procedure has been run on them.

These data are assigned morphological grammatemes (the full set of values) and syntactic grammatemes, and the nodes within the trees are reordered according to topic-focus articulation.

4.1 Differences from the Automatic Procedure

Differences in assignment of morphological grammatemes can be illustrated on several examples of verbal attributes:

- attribute Iterativeness can be set to IT1 (*he used to play tennis every day*);
- the interpretation of the tense is preferred to morphological realization (*he said he would come.*POST *on Monday*);
- attribute SENTMOD can have other values (*Come!.*IMPER; *Do you want to come?.*INTER) that cannot be assigned automatically according to the punctuation marks because of the presence of finite verbs within relative clauses (*he asked.*ENUNC *whether we could come.*INTER - interrogative mode is based on the lexical semantics of the finite verb).

The assignment of syntactic grammatemes is done only manually. The grammatemes specify the semantic interpretation mainly of temporal and locative functors. This interpretation is closely connected with the form of preposition, but instead of the original form of a preposition it contains a more general value. This set partly bears Czech forms of prepositions.

The assignment of syntactic grammatemes is related to some kind of "neutral" speech situation; this means for example that the sets for different locative functors are the same (*he was at school* vs. *he run to the school* with the same value of the syntactic grammateme).

The list of the most frequent values of syntactic grammateme GRAM:

- basic values (applicable for all functors)
  - **NIL** - unmarked realization
  - **APPX** - approximate value; *it costs about $ 100.*APPX
- values specific for locative and partly temporal functors
  - **v** - *he was in the garden.*LOC_v; *he was at the school.*LOC_v; *he runs to the cinema.*DIR3_v
  - **mezi.1** - *among*, *amid*
  - **mezi.2** - *between*
  - **na** - *knock at the door.*DIR3_na
  - **za** - *behind*
  - **vedle** - *beside*
  - **před** - *in_front_of*
- values only for temporal functors TWHEN and THO
  - **BEF** - *he arrived after the holiday.*BEF
  - **AFT** - *he was there a week ago.*AFT
- other values (used only with relevant functors)
  - functor=BEN (benefactor)
    - **NIL** - *for somebody*
    - **AGST** - *against somebody*

file	# sentences	# words and punct. marks	# tgts nodes	# incorrectly attached nodes	# incorrectly assigned functors
wsj_1789	48	1201	835	46 (5.5%)	143 (17.1%)
wsj_1790	45	1037	762	39 (5.1%)	122 (16.0%)
wsj_1795	60	1551	996	60 (6.0%)	191 (19.1%)
wsj_2100	53	1429	1002	94 (9.4%)	161 (16.0%)
wsj_2104	39	1298	889	49 (5.5%)	201 (22.6%)
total	245	6516	4484	288 (6.4%)	818 (18.2%)

Table 1: Evaluation of the quality of automatically created tectogrammatical trees (differences in deep word order are not counted here).

- functor=ACMP (accompaniment)
  - **NIL** - *with somebody*
  - **WOUT** - *without somebody*
- functor=CPR (comparison)
  - **NIL** - *the economy has become open as the other industrialized nations.*NIL
  - **DFR** - *he is more clever than me.*DFR
- functor=EXT
  - **MORE** - *he is too.*MORE *young to be her brother*
  - **LESS** - *she is almost.*LESS *thirty*
- functor=REG
  - **NIL** - *an excursus of little relevance to its central point.*NIL
  - **WOUT** - *no matter why he couldn't come.*WOUT *they...*

4.2 Evaluation

The evaluation of the automatic procedure is based on a comparison of the automatically generated and then manually corrected sentences from 5 files. The results are summarized in Table 1. Due to the low variation of the error rates, the presented transformation tool seems to be sufficiently robust.

5 Open questions

There are several unsolved topics in the automatic transformation of context-free trees to TGTS. The following three of them seem to us to be the most important:

- **assignment of functors and grammatemes**: finding rules for a better assignment of functors is needed for an automatic assignment of syntactic grammatemes;

- **morphological grammatemes for English**: creating a set of morphological grammatemes specific for English is necessary; solution should not be independent from the revision of the set of functors; for example *one of the most important topics* would be assigned in Czech the functor DIR1 because of the specific Czech surface realization (literally: *one from ...*): the question is whether we should use this functor also for the English variant or whether we should use a different, more general (or more specific?) functor, e.g. for selection from a semantic container or group;

- **topic-focus articulation**: the transformation tool described here doesn't even attempt at solving this problem because of its complexity; possible hints are definite and indefinite articles, information of verbal aspect and A'-movement traces in the original Penn-Treebank data.

6 Remarks on Text Generation

When generating the text from the tectogrammatical trees ([3]), the following difficult problems will have to be faced: how to (i) reconstruct function words, (ii) find an appropriate word order, and (iii) find an appropriate word form for each node.

6.1 Reconstructing function words

- **prepositions** - this is the most difficult problem; the majority of them could be reconstructed using the functor (if there is a dominating surface realization of the functor). However, some functors have a great variation in surface realization, none of them being significantly dominant; it is mainly the case of local and temporal circumstantials, for example in "*The cat slept on/below/behind/near the table*" the functor of "the table" is always LOC. The subtle differences should have been captured via grammatemes, but these are not assigned by the automatic procedure. In other words, when generating from the automatically generated trees, sometimes it is not possible to reconstruct the correct (pre)position of the sleeping cat (without looking into the FW attributes, which is a little bit of cheating).

- **auxiliary verbs** - the auxiliary verbs in complex verb forms can be derived from the combination of the values of the grammatemes ASPECT, SENTMOD, VERBMOD, TENSE (difficult, but feasible); note that negation is not a grammateme, but a child node of the verb node.

- **modal verbs** - grammateme DEONTMOD can be used (DEB → "must", HRT → "should", etc.).

- **subordinating conjunctions** - this should be quite straightforward: a table which maps the functor of the head of the subordinating clause to the appropriate conjunction (COND → "if" etc.) could be hopefully constructed.

- **determiners** - the "official" tectogrammatical level does not give a tool for representing the determiners (it was developed for Czech, which does not have them). It is obvious that in an English sentence the determiners cannot be first deleted and then reconstructed with certainty without knowing the numerous conventions, the world, and—what is the worst—the sentence context. However, after an appropriate study, at least the topic-focus annotation and the deep word order could be used for inserting determiners.

Besides function words, also the punctuation marks have to be reconstructed.

6.2 Finding appropriate word forms

Word forms[12] are not present on the tectogrammatical level and must be derived from the lemma and the values of respective grammatemes. This is trivial in some cases (e.g., generating

[12] Whenever we speak about finding word forms here, we mean in fact finding the appropriate POS tags, from which (and the lemma) the word forms can be easily created using any morphological dictionary of English.

plural for nouns is influenced only by the grammateme NUMBER of the same node), but it is non-trivial in others:

- **subject-verb agreement** - the correct verb form must agree in person and number with the subject (the subject might be coordinated).
- **complex verb forms** - several grammatemes (TENSE, ASPECT, SENTMOD, VERBMOD, DEONTMOD, NUMBER, PERSON), and the existence of the negation child node must be considered when searching for the correct word forms of a given autosemantic verb and possibly auxiliary verb(s).

7 Conclusion

We have shown that the phrase trees from the Penn Treebank can be automatically transformed into the tectogrammatical trees with a reasonably high reliability. The quality evaluation (based on the comparison with manually annotated trees) can be summarized as follows: there are about 6% of wrongly aimed dependecies (wrongly attached nodes), and about 18% of wrongly assigned functors.

Acknowledgment

Our thanks go to Jan Hajič, Eva Hajičová, Jiří Havelka, and Petr Sgall for many useful comments on this text.

References

[1] Bies, Ann, Mark Fergusona, Karen Katz, and Robert MacIntyre. Bracketing Guidelines for Treebank II Style. Penn Treebank Project, University of Pennsylvania (1995)

[2] Eisner, Jason: Smoothing a Probabilistic Lexicon Via Syntactic Transformations. University of Pennsylvania (2001)

[3] Hajič Jan et al.: Generation in the context of MT. Final Report. Center for Language and Speech Processing, Johns Hopkins University, Baltimore, MD. In prep. (2002)

[4] Hajič, Jan, Petr Pajas, Barbora Hladká: The Prague Dependency Treebank: Annotation Structure and Support, IRCS Workshop on Linguistic Databases, Philadelphia, PA (2001)

[5] Hajičová, Eva, Jan Hajič, Barbora Hladká, Martine Holub, Petr Pajas, Veronika Řezníčková, and Petr Sgall: The Current Status of the Prague Dependency Treebank, proceedings of Text, Speech and Dialogue, Springer-Verlag (2001)

[6] Mitchell P. Marcus, Beatrice Santorini, Mary Ann Marcinkiewicz: Building a Large Annotated Corpus of English: The Penn Treebank, Computational Linguistics (1994)

[7] Sgall, Petr, Eva Hajičová, and Jarmila Panevová: The Meaning of the Sentence and Its Semantic and Pragmatic Aspects. Academia/Reidel Publishing Company, Prague, Czech Republic/Dordrecht, Netherlands (1986)

A Notation used in the Penn Treebank

The following summary was extracted from [1].

A.1 Part-of-Speech Tags

CC coordinating conjunction (*and*)
CD cardinal number (*1, third*)
DT determiner (*the*)
EX existential there (*there is*)
FW foreign word (*d'hoevre*)
IN preposition/subordinating conjunction (*in, of, like*)
JJ adjective (*green*)
JJR adjective, comparative (*greener*)
JJS adjective, superlative (*greenest*)
LS list marker (*1)*)
MD modal (*could, will*)
NN noun, singular or mass (*table*)
NNS noun plural (*tables*)
NNP proper noun, singular (*John*)
NNPS proper noun, plural (*Vikings*)
PDT predeterminer (*¡i¿both¡/i¿ the boys*)
POS possessive ending (*friend's*)
PRP personal pronoun (*I, he, it*)
PRP$ possessive pronoun (*my, his*)
RB adverb (*however, usually, naturally, here, good*)
RBR adverb, comparative (*better*)
RBS adverb, superlative (*best*)
RP particle (*give up*)
TO to (*to go, to him*)
UH interjection (*uhhuhhuhh*)
VB verb, base form (*take*)
VBD verb, past tense (*took*)
VBG verb, gerund/present participle (*taking*)
VBN verb, past participle (*taken*)
VBP verb, sing. present, non-3d (*take*)
VBZ verb, 3rd person sing. present (*takes*)
WDT wh-determiner (*which*)
WP wh-pronoun (*who, what*)
WP$ possessive wh-pronoun (*whose*)
WRB wh-abverb (*where, when*)

A.2 Phrase Labels

S simple declarative clause
SBAR clause introduced by a subord. conjunction
SBARQ direct question introduced by a wh-word
SINV inverted declarative sentence
SQ inverted yes/no question
ADJP adjective phrase
ADVP adverb phrase
CONJP conjunction phrase
FRAG fragment
INTJ interjection
LST list marker
NAC not a constituent
NX something like N-bar level
PP prepositional phrase
PRN parenthetical
PRT particle
QP quantifier phrase
RRC reduced relative clause
UCP unlike coordinated phrase
VP verb phrase
WHADJP wh-adjective phrase
WHADVP wh-adverb phrase
WHNP wh-noun phrase
WHPP wh-prepositional phrase
X unknown

A.3 Function tags

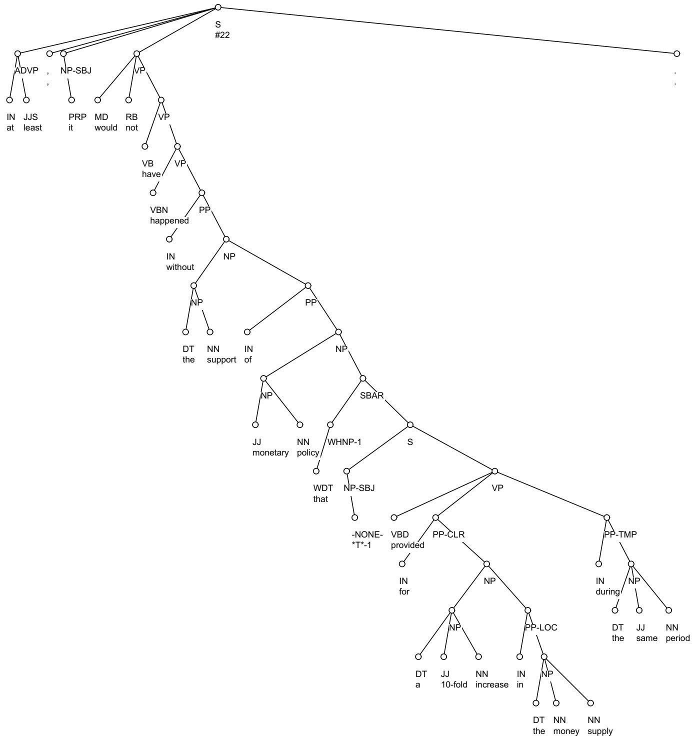
-ADV adverbial
-NOM nominal
-DTV dative
-LGS logical subject
-PRD predicate
-PUT loc. complement of put
-SBJ surface subject
-TPC topicalized
-VOC vocative
-BNF benefactive
-DIR direction
-EXT extent
-LOC locative
-MNR manner
-PRP purpose or reason
-TMP temporal
-CLR closely related
-CLF cleft
-HLN headline
-TTL title

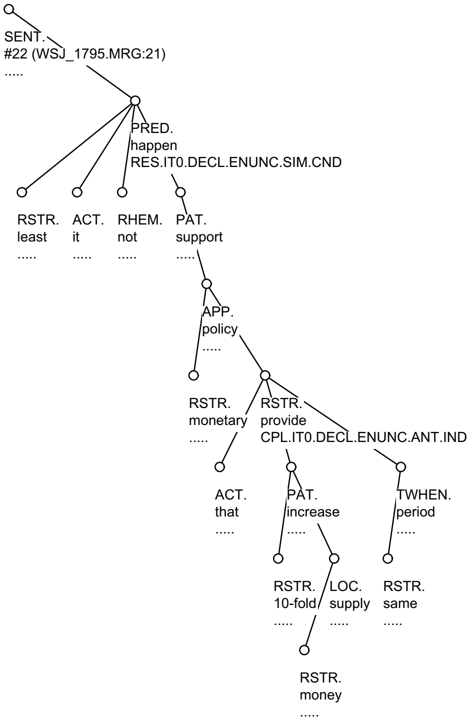
B Alphabetically Ordered List of 40 Functors Most Frequent in the Prague Dependency Treebank

ACMP (accompaniment): mothers with *children*
ACT (actor): *Peter* read a letter.
ADDR (addressee): Peter gave *Mary* a book.
ADVS (adversative): He came there, *but* didn't stay long.
AIM (aim): He came there to *look* for Jane.
APP (appurtenance, i.e., possession in a broader sense): *John's* desk
APPS (apposition): Charles the Fourth, (i.e.) *the Emperor*
ATT (attitude): They were here *willingly.*
BEN (benefactive): She made this for her *children.*
CAUS (cause): She did so since they *wanted* it.
COMPL (complement): They painted the wall *blue.*
COND (condition):If they *come* here, we'll be glad.
CONJ (conjunction): Jim *and* Jack
CPR (comparison): *taller* than Jack
CRIT (criterion): According to *Jim*, it was raining there.
DENOM (denomination): *Chapter 5* (e.g. as a title)
DIFF (difference): taller by two *inches*
DIR1 (direction-from): He went from the *forest* to the village.
DIR2 (direction-through): He went through the *forest* to the village
DIR3 (direction-to): He went from the forest to the *village.*
DISJ (disjunction): here *or* there
DPHR (dependent part of a phraseme): in *no* way, *grammar* school
EFF (effect): We made him the *secretary.*
EXT (extent): *highly* efficient
FPHR (foreign phrase): *dolcissimo*, as they say
ID (entity): the river *Thames*
LOC (locative): in *Italy*
MANN (manner): They did it *quickly.*
MAT (material): a bottle of *milk*
MEANS (means): He wrote it by *hand.*
MOD (mod): He *certainly* has done it.
PAR (parentheses): He has, as we *know*, done it yesterday.
PAT (patient): I saw *him.*
PHR (phraseme): in no *way*, grammar *school*
PREC (preceding, particle referring to context): *therefore*, *how ever*
PRED (predicate): I *saw* him.
REG (regard): with regard to *George*
RHEM (rhematizer, focus sensitive particle): *only*, *even*, *also*
RSTR (restrictive adjunct): a *rich* family
THL (temporal-how-long): We were there for three *weeks.*
THO (temporal-how-often) We were there very *often.*
TWHEN (temporal-when): We were there at *noon.*

C Samples of WSJ phrase trees and their (automatically created) tectogrammatical counterparts

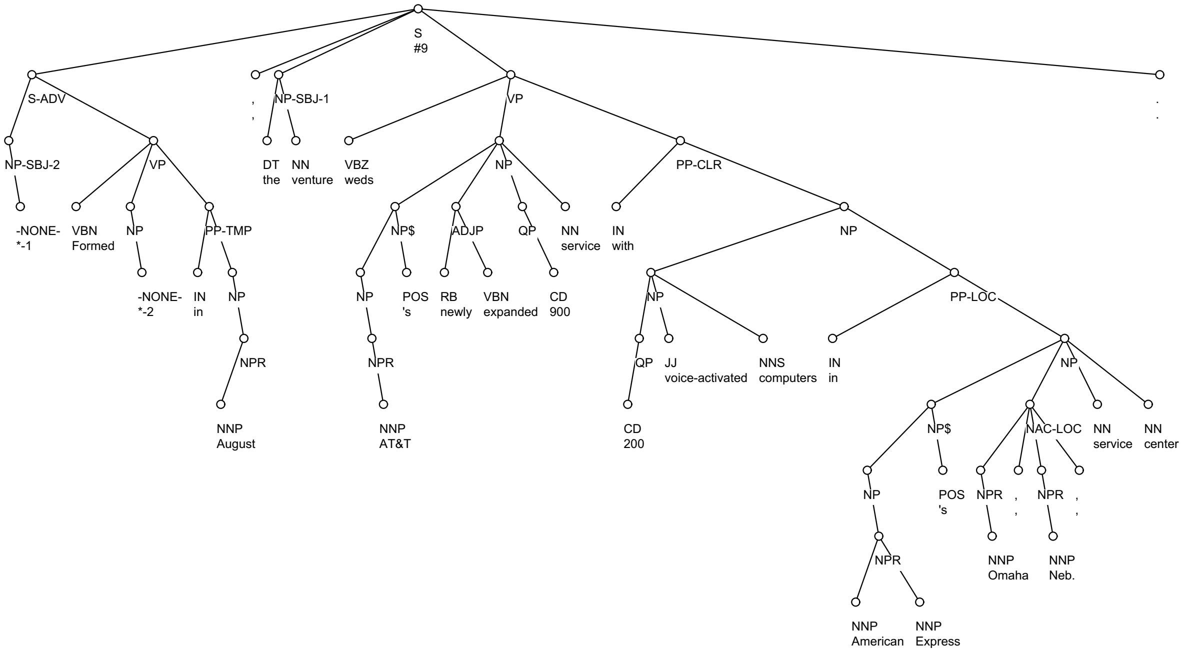
C.1 Sample sentence: *"At least, it would not have happened without the support of monetary policy that provided for a 10-fold increase in the money supply during the same period."*

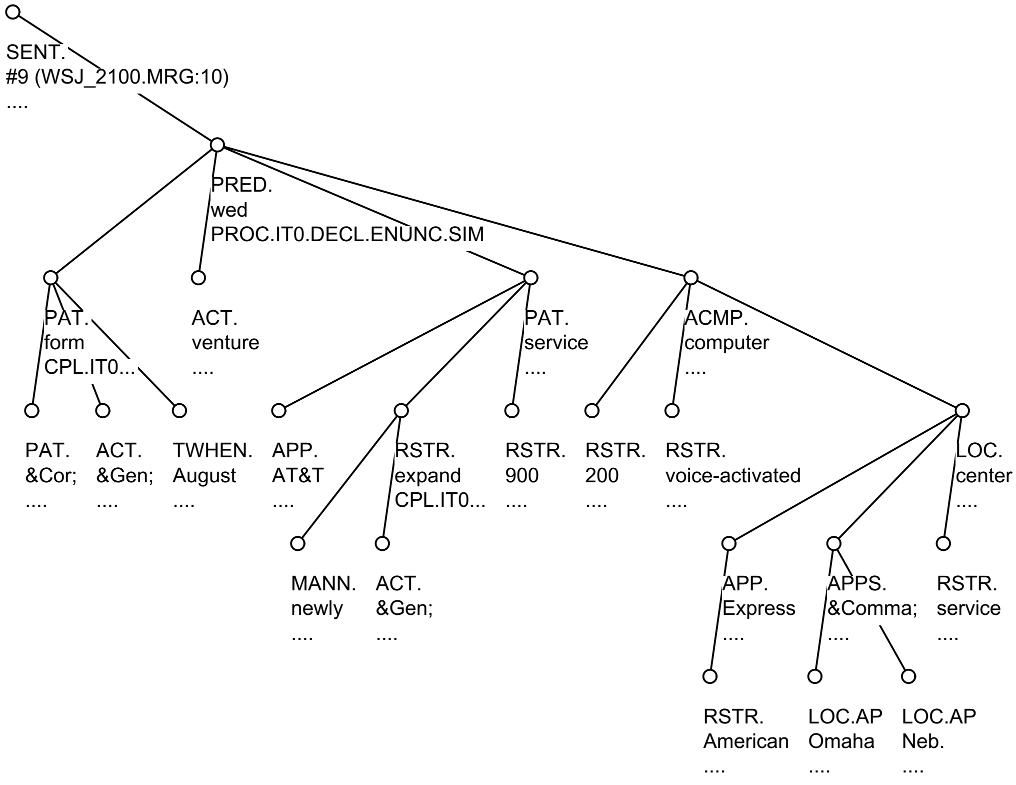




C.2 Sample sentence: *"Formed in August, the venture weds AT&T's newly expanded 900 service with 200 voice-activated computers in American Express's Omaha, Neb., service center."*

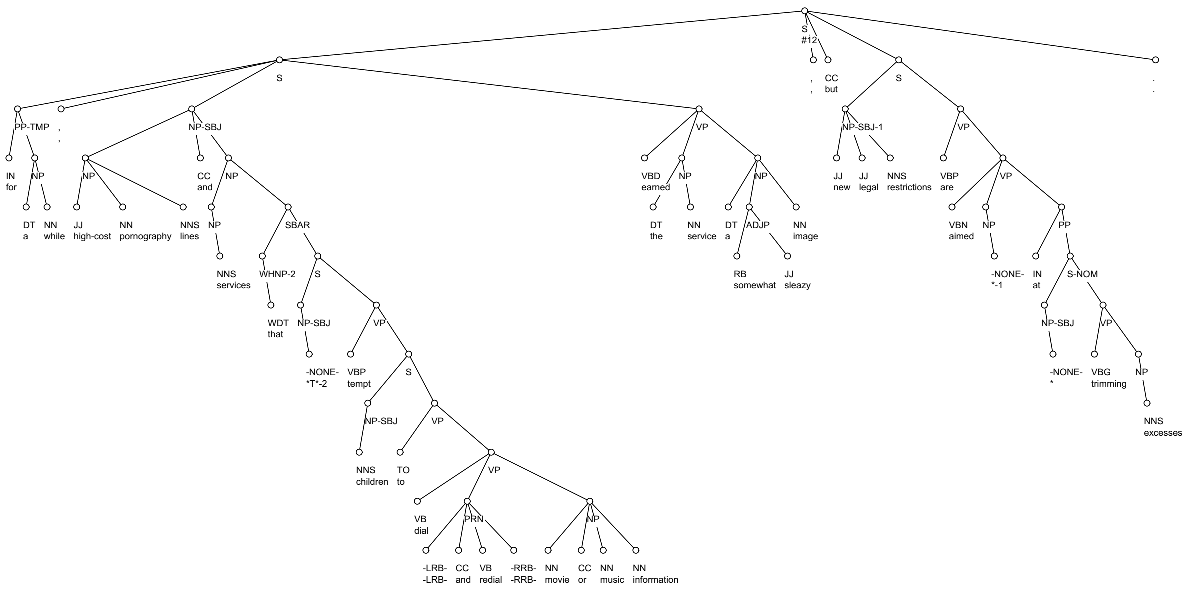
Note the A-movement traces and the apposition (the grammateme MEMBEROF is filled).

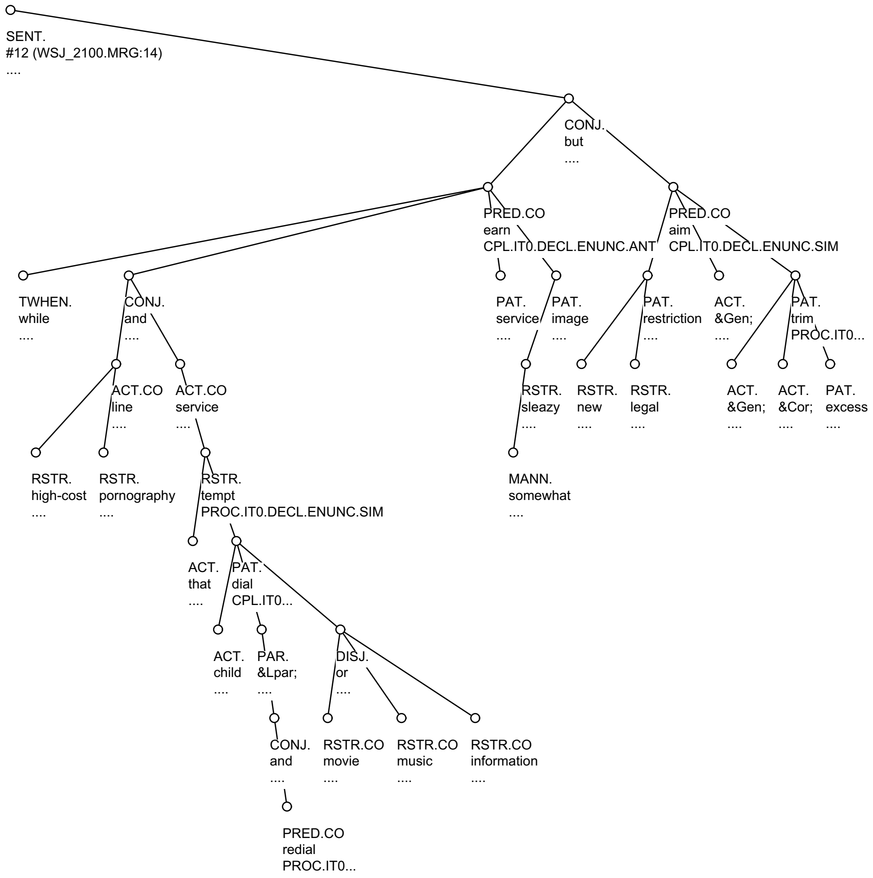




C.3 Sample sentence: *"For a while, high-cost pornography lines and services that tempt children to dial (and redial) movie or music information earned the service a somewhat sleazy image, but new legal restrictions are aimed at trimming excesses."*

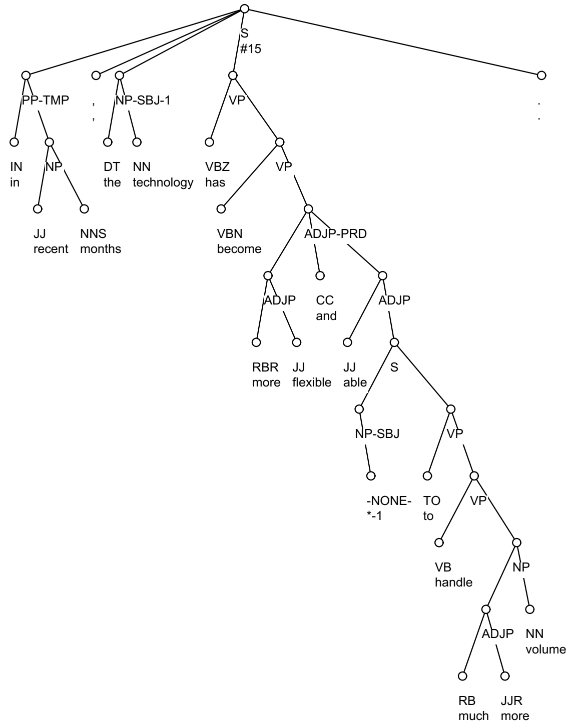
Note how the A-movement traces were processed (for passive voice). The figure also contains the coordination (again, the grammateme MEMBEROF is assigned) and the parenthesis.

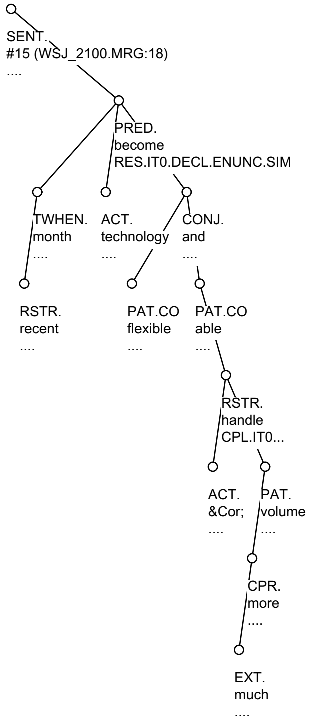




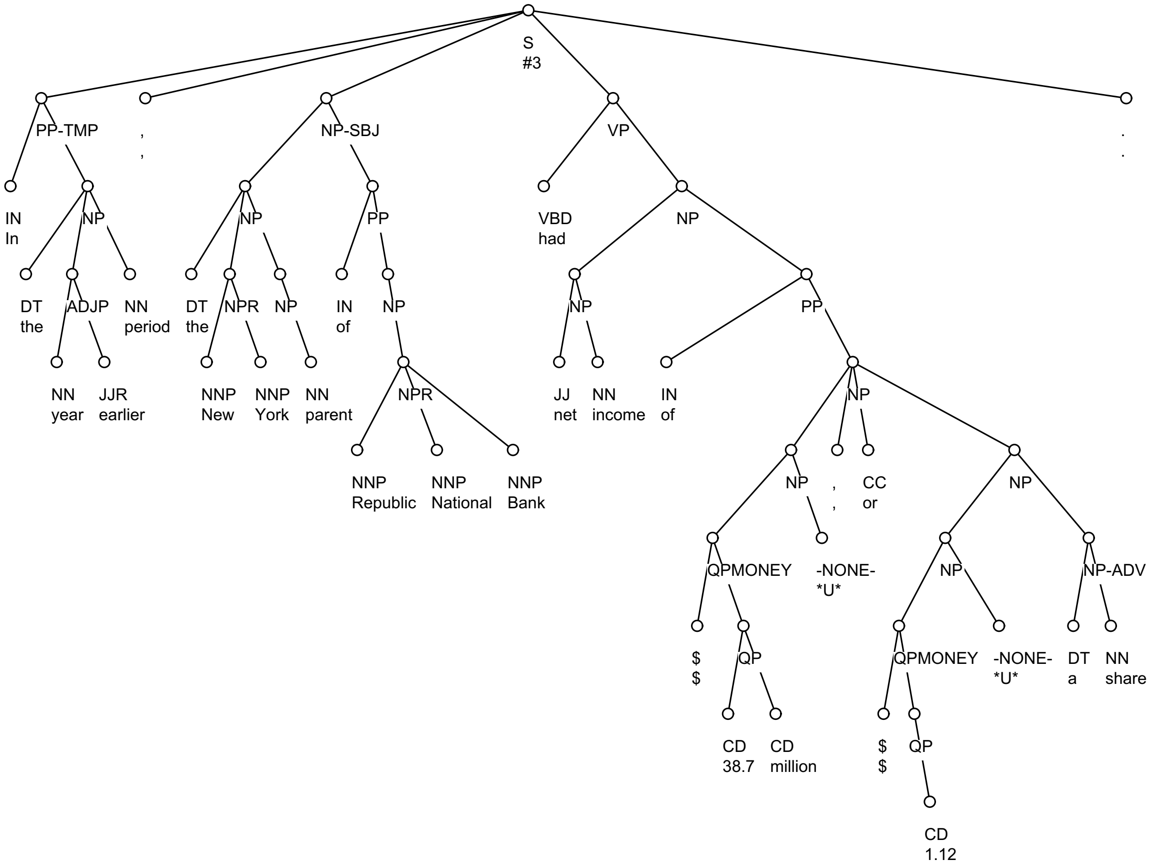
C.4 Sample sentence: *"In recent months, the technology has become more flexible and able to handle much more volume."*

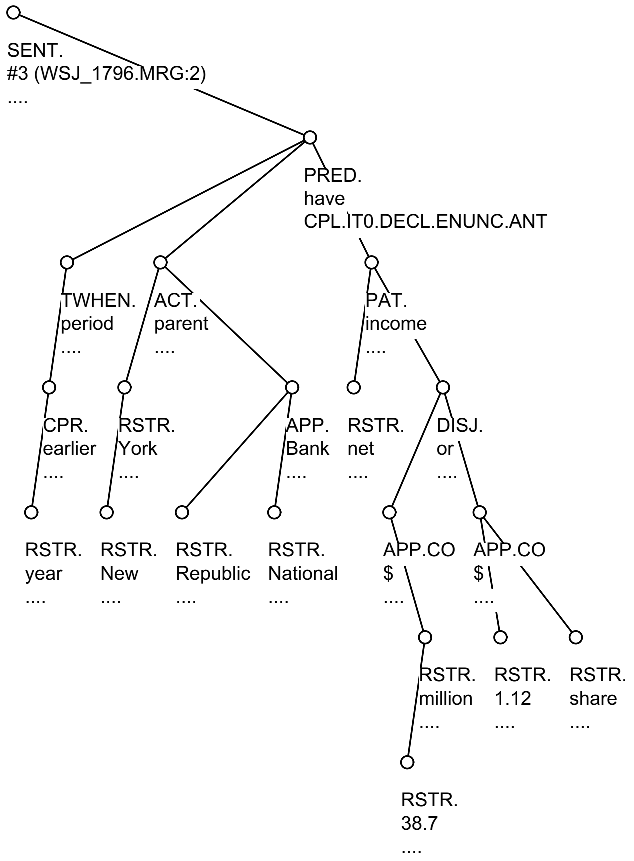
Note the comparative form of adjective *flexible*. The grammateme DEGCMP of the corresponding node is set to COMP.





C.5 Sample sentence: *"In the year earlier period, the New York parent of Republic National Bank had net income of $ 38.7 million, or $ 1.12 a share."*


S
#3
PP-TMP
,
,
NP-SBJ
VP
.
.
IN
In
NP
DT
the
ADJP
NN
period
NN
year
JJR
earlier
NP
DT
the
NPR
NP
NNP
New
NNP
York
NN
parent
PP
IN
of
NP
NPR
NNP
Republic
NNP
National
NNP
Bank
VBD
had
NP
NP
JJ
net
NN
income
PP
IN
of
NP
NP
,
,
CC
or
NP
QPMONEY
-NONE-
U
$
$
QP
CD
38.7
CD
million
NP
NP-ADV
QPMONEY
-NONE-
U
DT
a
NN
share
$
$
QP
CD
1.12


SENT.
#3 (WSJ_1796.MRG:2)
....
PRED.
have
CPL.IT0.DECL.ENUNC.ANT
TWHEN.
period
....
ACT.
parent
....
PAT.
income
....
CPR.
earlier
....
RSTR.
York
....
APP.
Bank
....
RSTR.
net
....
DISJ.
or
....
RSTR.
year
....
RSTR.
New
....
RSTR.
Republic
....
RSTR.
National
....
APP.CO
$
....
APP.CO
$
....
RSTR.
million
....
RSTR.
1.12
....
RSTR.
share
....
RSTR.
38.7
....