Dependency Syntax in the CoNLL Shared Task 2008

Richard Johansson
Department of Computer Science, Lund University
richard@cs.lth.se

Draft version 0.5, March 10, 2007

1 Introduction

This document gives a brief overview of the conversion of the Penn Treebank (Marcus et al., 1993, 1994) to the dependency structures used in the CoNLL-2008 Shared Task. Our dependency framework has the following properties:

- *single-head*: every word has exactly one parent, except the root, which has no parent.
- *single-root*: only one word in the sentence is root.
- *traceless*: the dependency structures use no empty categories. Special arc labels are used to encode gapping.
- *nonprojective*: some long-distance syntactic phenomena are represented in the dependency structure by means of non-local links.

The conversion procedure relies on earlier work on constituent-to-dependency conversion (Magerman, 1994; Collins, 1999; Yamada and Matsumoto, 2003; Johansson and Nugues, 2007). In addition, we imported dependencies inside NPs and hyphenated words from a version of the Penn Treebank mapped into GLARF, the Grammatical and Logical Argument Representation Framework (Meyers et al., 2001).

To assign dependency labels, we used the following general principles:

- If there is a Treebank label other than `CLR`, `HLN`, `NOM`, `TPC`, or `TTL`: use this label.
- If the link is inside an NP or a hyphenated word: use the label from GLARF.
- Else infer a label using a set of rules.

The complete set of labels is listed in Section 4.

2 Head Percolation Rules

Following earlier work on conversion from constituents to dependencies, the central principle of the conversion procedure is to assign a head word to each constituent. The head words are found by *head percolation*: recursively searching through constituents using heuristic rules to determine in which child the head can be found. When a head word has been assigned to each constituent, conversion to dependency structure is straightforward.

The following table lists the head percolation rules for each phrase type. The second column indicates search direction, and the third is a priority list of phrase types to look for. For instance, to find the head of an S phrase, we look from right to left for a VP. If no VP was found, look for anything with a PRD function tag, and so on.

<table>
<tr><td>ADJP</td><td>←</td><td>NNS QP NN $ ADVP JJ VBN VBG ADJP JJR NP JJS DT FW RBR RBS SBAR RB</td></tr>
<tr><td>ADVP</td><td>→</td><td>RB RBR RBS FW ADVP TO CD JJR JJ IN NP JJS NN</td></tr>
<tr><td>CONJP</td><td>→</td><td>CC RB IN</td></tr>
<tr><td>FRAG</td><td>→</td><td>(NN⋆ | NP) W⋆ SBAR (PP | IN) (ADJP | JJ) ADVP RB</td></tr>
<tr><td>INTJ</td><td>←</td><td>⋆</td></tr>
<tr><td>LST</td><td>→</td><td>LS :</td></tr>
<tr><td>NAC</td><td>←</td><td>NN⋆ NP NAC EX $ CD QP PRP VBG JJ JJS JJR ADJP FW</td></tr>
<tr><td>NP, NX, WHNP</td><td>←</td><td>(NN⋆ | NX) NP-ε JJR CD JJ JJS RB QP NP</td></tr>
<tr><td>PP, WHPP</td><td>→</td><td>IN TO VBG VBN RP FW</td></tr>
<tr><td>PRN</td><td>→</td><td>S⋆ N⋆ W⋆ PP|IN ADJP|JJ⋆ ADVP|RB⋆</td></tr>
<tr><td>PRT</td><td>→</td><td>RP</td></tr>
<tr><td>QP</td><td>←</td><td>$ IN NNS NN JJ RB DT CD NCD QP JJR JJS</td></tr>
<tr><td>RRC</td><td>→</td><td>VP NP ADVP ADJP PP</td></tr>
<tr><td>S</td><td>←</td><td>VP ⋆-PRD S SBAR ADJP UCP NP</td></tr>
<tr><td>SBAR</td><td>←</td><td>S SQ SINV SBAR FRAG IN DT</td></tr>
<tr><td>SBARQ</td><td>←</td><td>SQ S SINV SBARQ FRAG</td></tr>
<tr><td>SINV</td><td>←</td><td>VBZ VBD VBP VB MD VP ⋆-PRD S SINV ADJP NP</td></tr>
<tr><td>SQ</td><td>←</td><td>VBZ VBD VBP VB MD ⋆-PRD VP SQ</td></tr>
<tr><td>UCP</td><td>→</td><td>⋆</td></tr>
<tr><td>VP</td><td>→</td><td>VBD VBN MD VBZ VB VBG VBP VP ⋆-PRD ADJP NN NNS NP</td></tr>
<tr><td>WHADJP</td><td>←</td><td>CC WRB JJ ADJP</td></tr>
<tr><td>WHADVP</td><td>→</td><td>CC WRB</td></tr>
<tr><td>X</td><td>→</td><td>⋆</td></tr>
</table>

3 Treatment of Some Complex Linguistic Phenomena

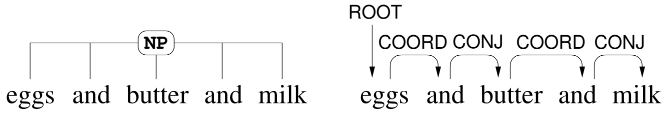
This section lists a number of non-trivial constructions for which attachment is determined by special heuristics rather than head percolation.

3.1 Coordination

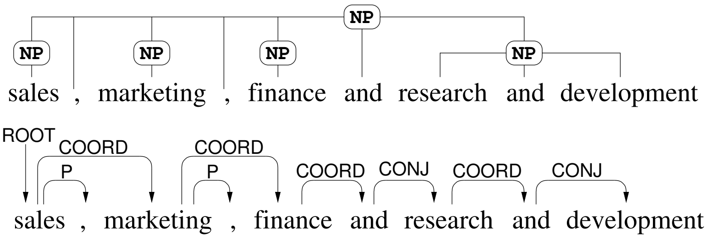
We use Mel'čuk-style analysis of coordination (Mel'čuk, 1988). Specifically, we use the following conventions:

- The first conjunct is regarded as the head of the coordinated structure.
- The second conjunct is linked to the first via a COORD link.
- If a coordinating conjunction is present, it becomes the head of the second conjunct using a CONJ link.

- Coordination is right-associative.

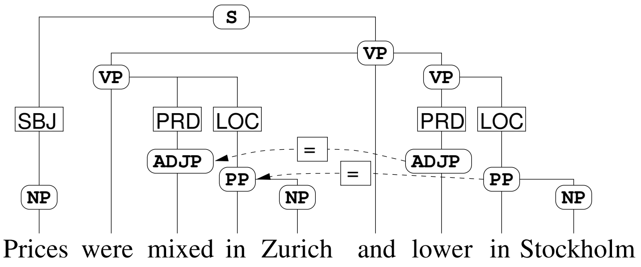


A more complex example involving a multi-level coordination with commas is shown below.

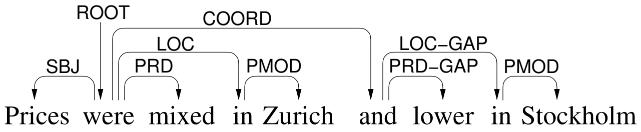

NP
NP NP NP NP
sales , marketing , finance and research and development
ROOT
COORD
P
COORD
P
COORD CONJ COORD CONJ
sales , marketing , finance and research and development

3.2 Gapping

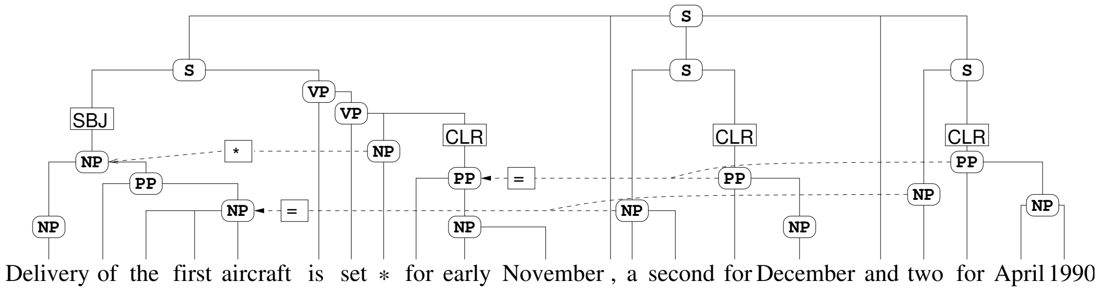
Gapping refers to the phenomenon that the head of the second conjunct in a coordination is dropped, such as the verb in this example:

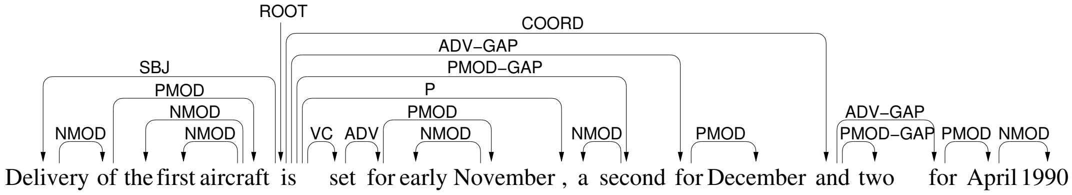


To handle this phenomenon without introducing empty categories, we follow the analysis in the Danish Dependency Treebank (Trautner Kromann et al., 2004), meaning that the parts of the second conjunct are attached to the conjunction, and the links carry a GAP label. If there is no conjunction, the parts are attached to the head of the first conjunct.



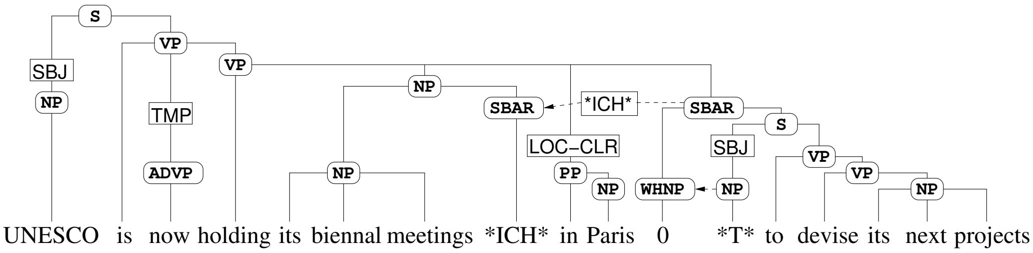
Here is a more complex example:

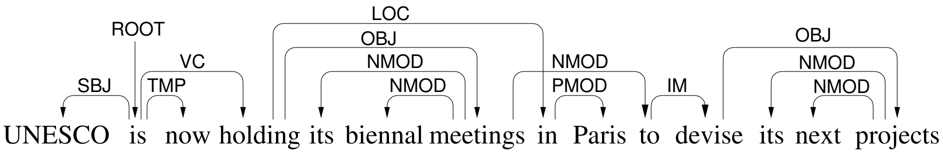

S
S
S
S
VP
VP
SBJ
*
NP
NP
CLR
CLR
CLR
PP
=
PP
PP
PP
NP
=
NP
NP
NP
NP
NP
NP
Delivery of the first aircraft is set * for early November , a second for December and two for April 1990


ROOT
COORD
ADV−GAP
PMOD−GAP
SBJ
PMOD
P
NMOD
PMOD
NMOD
NMOD
VC
ADV
NMOD
NMOD
PMOD
ADV−GAP
PMOD−GAP
PMOD
NMOD
Delivery of the first aircraft is set for early November , a second for December and two for April 1990

3.3 Discontinuous Structures (*ICH*)

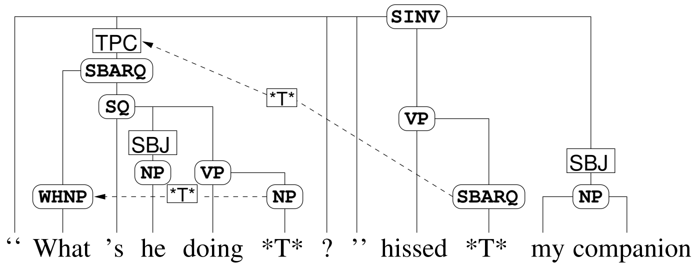
Discontinuous structures, represented in the Treebank by the *ICH* link ("interpret constituent here"), are simply handled by introducing nonprojective links.

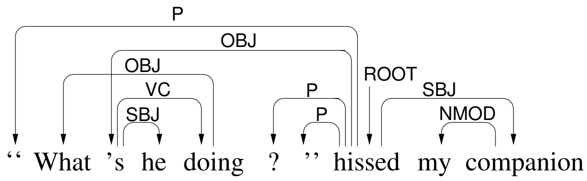

S
VP
VP
SBJ
NP
TMP
ADVP
NP
NP
SBAR
ICH
SBAR
S
LOC−CLR
PP
NP
SBJ
WHNP
NP
VP
VP
NP
UNESCO is now holding its biennal meetings *ICH* in Paris 0 *T* to devise its next projects


ROOT
LOC
OBJ
VC
NMOD
SBJ
TMP
NMOD
NMOD
PMOD
IM
OBJ
NMOD
NMOD
UNESCO is now holding its biennal meetings in Paris to devise its next projects

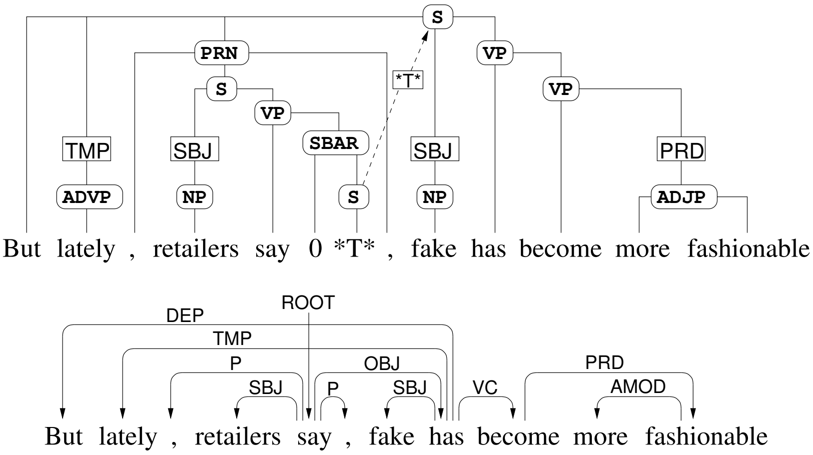
3.4 Nonlocal Dependencies

Some phenomena such as *wh*-movement and topicalization are represented in the Treebank by the empty category *T* ("trace"). When building the dependency representation, these links take priority over constituent attachment.



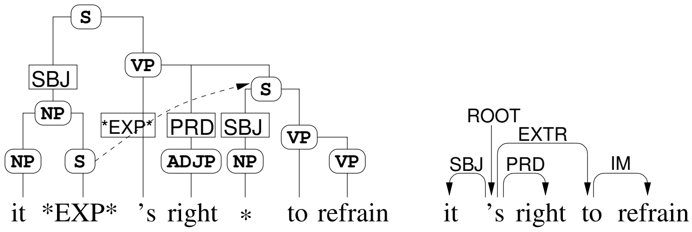


For some discontinuous structures, the Treebank uses a parenthetical and a trace linking the parenthetical to the top node. In these cases, the conversion procedure moves the parenthetical over its parent to break the cycle.


S
PRN
VP
S
VP
T
VP
SBAR
TMP
SBJ
SBJ
PRD
ADVP
NP
S
NP
ADJP
But lately , retailers say 0 *T* , fake has become more fashionable
ROOT
DEP
TMP
P
OBJ
SBJ
P
SBJ
VC
PRD
AMOD
But lately , retailers say , fake has become more fashionable

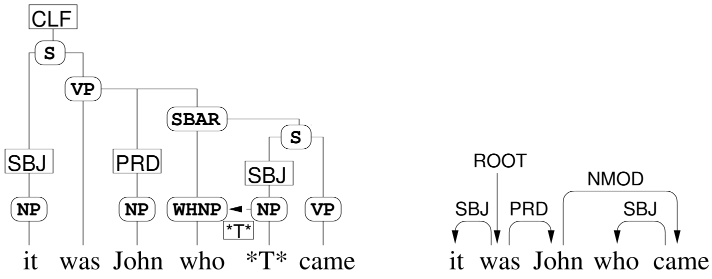
3.5 Expletive *it*

The extraposed element in an expletive construction gets an EXTR label.



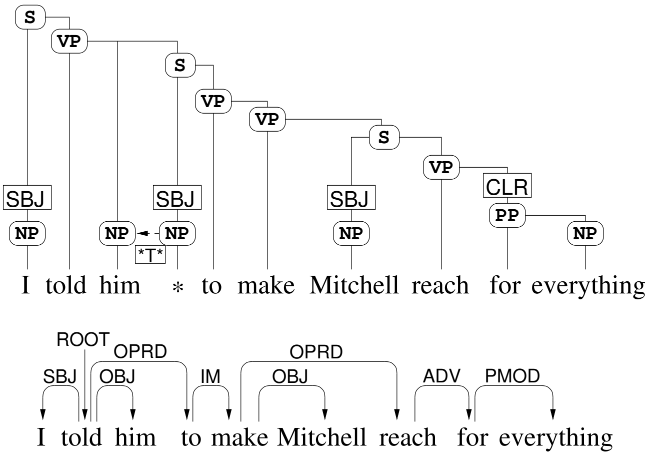
3.6 Cleft Sentences

The extraposed element in an *it*-cleft is treated as if attached to its antecedent.

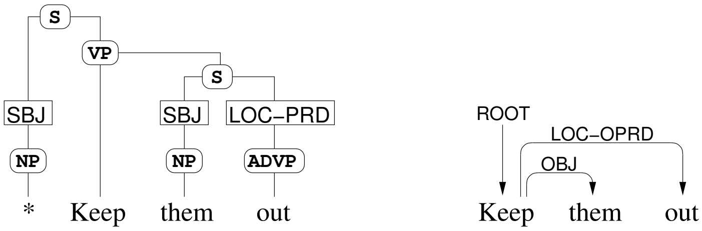


3.7 Object Complements

Some constructions, such as "small clauses" (see Bies et al. (1995), section 15), are represented in the Treebank using an S node directly inside a verb phrase. Since the case and position of the "subject" of such an S is determined by the voice of the verb in the enclosing VP, we move it to the object position in the VP. The remainder of the S clause is labeled with an OPRD label ("object predicative complement").

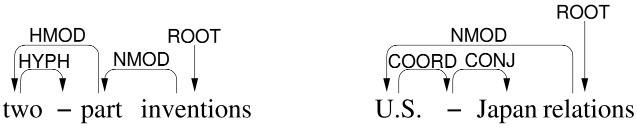

S
VP
S
VP
VP
S
VP
SBJ
SBJ
SBJ
CLR
NP
NP
NP
T
NP
PP
NP
I told him * to make Mitchell reach for everything
ROOT
OPRD
OPRD
SBJ
OBJ
IM
OBJ
ADV
PMOD
I told him to make Mitchell reach for everything

If the head part of the `S` has a function label, and this label is not `PRD`, this label is prefixed to the `OPRD` label, as in the example below.



3.8 Hyphenated Structures

To represent hyphenated structures, we have introduced two new POS tags: `HYPH` for hyphens and `PRF` for prefixes such as *non-* and *anti-*. We distinguish two types of relations inside hyphenated words: modification and coordination. The following figure shows how they are represented.



4 List of Dependency Relations

Note that labels may be combined, such as `LOC−OPRD` or `PRD−GAP`.

4.1 Labels Retained from the Penn Treebank

Label	Meaning
ADV	Unclassified adverbial
BNF	Benefactor (the *for* phrase for verbs that undergo dative shift)
DIR	Direction
DTV	Dative (the *to* phrase for verbs that undergo dative shift)
EXT	Extent
LGS	Logical subject
LOC	Location
MNR	Manner
PRD	Predicative complement
PRP	Purpose or reason
PUT	Various locative complements of the verb *put*
SBJ	Subject
TMP	Temporal
VOC	Vocative

4.2 Labels Derived from GLARF

Label	Meaning
APPO	Apposition
HMOD	Modifier in hyphenation, such as *two* in *two-part*
HYPH	Between first part of hyphenation and hyphen
NAME	Name-internal link
POSTHON	Posthonorifics such as *Jr, Inc.*
SUFFIX	Possessive *'s*
TITLE	Titles such as *Mr, Dr*

4.3 Inferred Labels

Label	Meaning
AMOD	Modifier of adjective or adverb
CONJ	Between conjunction and second conjunct in a coordination
COORD	Coordination
DEP	Unclassified relation
EXTR	Extraposed element in expletive constructions
GAP	Gapping: between conjunction and the parts of a structure with an ellipsed head
IM	Between infinitive marker and verb
NMOD	Modifier of nominal
OBJ	Direct or indirect object or clause complement
OPRD	Object complement
P	Punctuation
PMOD	Between preposition and its child in a PP
PRN	Parenthetical
PRT	Particle
ROOT	Root
SUB	Between subordinating conjunction and verb
VC	Verb chain

References

Bies, A., Ferguson, M., Katz, K., MacIntyre, R., Tredinnick, V., Kim, G., Marcinkiewicz, M., and Schasberger, B. (1995). Bracketing guidelines for Treebank II style Penn Treebank project.

Collins, M. (1999). Head-driven statistical models for natural language parsing. Ph.D. thesis, University of Pennsylvania, Philadelphia.

Johansson, R. and Nugues, P. (2007). Extended constituent-to-dependency conversion for English. In *Proceedings of NODALIDA 2007*.

Magerman, D. M. (1994). Natural language parsing as statistical pattern recognition. Ph.D. thesis, Stanford University.

Marcus, M., Kim, G., Marcinkiewicz, M., MacIntyre, R., Bies, A., Ferguson, M., Katz, K., and Schasberger, B. (1994). The Penn treebank: Annotating predicate argument structure. In *ARPA Human Language Technology Workshop*.

Marcus, M., Santorini, B., and Marcinkiewicz, M. (1993). Building a large annotated corpus of English: the Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Mel'čuk, I. A. (1988). *Dependency Syntax: Theory and Practice*. State University Press of New York.

Meyers, A., Grishman, R., Kosaka, M., and Zhao, S. (2001). Covering treebanks with GLARF. In *Proceedings of the ACL/EACL 2001 Workshop on Sharing Tools and Resources for Research and Education*.

Trautner Kromann, M., Mikkelsen, L., and Kern Lynge, S. (2004). Annotation guide for the Danish Dependency Treebank. Available on the web at `http://www.isv.cbs.dk/~mtk/treebank`.

Yamada, H. and Matsumoto, Y. (2003). Statistical dependency analysis with support vector machines. In *Proceedings of 8th International Workshop on Parsing Technologies*, pages 195–206.