

Dependency parser demo

Timo Järvinen and Pasi Tapanainen

University of Helsinki, Department of General Linguistics
Research Unit for Multilingual Language Technology
P.O. Box 4, FIN-00014 University of Helsinki, Finland
{Pasi.Tapanainen,Timo.Jarvinen}@ling.Helsinki.fi

1 Introduction

We are concerned with surface-syntactic parsing of running text. Our main goal is to describe a syntactic analysis of sentences using dependency links that show the head-dependent relations between words. The new dependency parser¹ (Tapanainen and Järvinen, 1997; Järvinen and Tapanainen, 1997) belongs to a continuous effort to apply rule-based methods to natural languages. It can be seen as a relative of the Constraint Grammar framework (Karlsson et al., 1995), for many features of the system have been derived from it. The syntactic description in the English Constraint Grammar (ENGCG) is implicitly dependency oriented; it contains tags for heads and modifiers but not explicit links between them (see Figure 2).

Although, the new syntactic formalism differs much from the Constraint Grammar's formalisms, the basic rule types of the older formalism have been preserved among the new ones. Also, the rules are independent, and they describe syntax in a piecemeal fashion. The new dependency parser creates explicit links between the elements of the sentence (in Figure 1) while still retaining the shallower representation similar to ENGCG (in Figure 2). The parser applies the ENGTWOL lexicon designed originally by Juha Heikkilä and Atro Voutilainen. Also, the reliable parts of the ENGCG's morphological disambiguator by Atro Voutilainen are applied.

The parser has been tested in Sun workstation and in PCs under Linux. The syntactic analysis is modest in time and space requirements: the size of the process (the syntactic analysis only) is less than 2 MB and it runs in a Pentium 90 MHz machine at the speed of 200 words per second. We have tested the parser on bigger texts to test its usability in corpus linguistic and lexicographic work. By now, some 30 million words have been parsed.

2 The dependency model

Our syntactic description can be seen as a formalisation of Tesnière's (1959) original dependency theory. The dependency model adopted to our description differs in various respects from the post-Tesnièrean development of dependency theory, though many of the features are recognised elsewhere. The main features of the parsing system and the adopted dependency theory are:

- The basic syntactic element is not a word, but a nucleus. This is related to the internal organisation of the grammar, though the default output shows the dependency links between surface words.
- Every element has one and only one head (uniqueness).
- The result is a tree.
- Functional dependencies are expressed by link names.
- The links may cross (non-projective constructions are allowed).
- Modifiers are not obligatory; valency defines possibility rather than obligatoriness to have arguments.
- The grammar is not generative. This means that the parser accepts every input sentence, and returns an analysis even for ungrammatical sentences to the extent the structure is recoverable.
- The dependency description is monostratal, i.e. there is one level of syntactic description, the surface-syntactic description, and no transformations.

¹<http://www.ling.helsinki.fi/~tapanain/dg/>

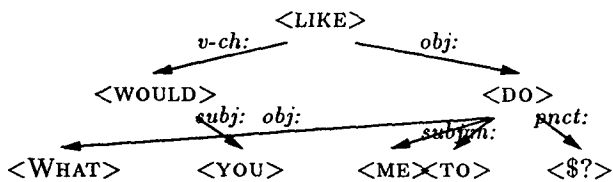


Figure 1: Dependency tree

```

<What>
  "what" <**CLB> PRON WH SG/PL @OBJ
<would>
  "would" V AUXMOD VFIN @+FAUXV
<you>
  "you" PRON PERS NOM SG2/PL2 @SUBJ
<like>
  "like" V INF @-FMAINV
<me>
  "i" PRON PERS ACC SG1 @OBJ
<to>
  "to" INFMARK @INFMARK
<do>
  "do" V INF @-FMAINV
<?>
  
```

Figure 2: ENGCG-style output

References

- Timo Järvinen and Pasi Tapanainen. 1997. A dependency parser for English. Technical Report TR-1, Department of General Linguistics, University of Helsinki, Finland, February.
- Fred Karlsson, Atro Voutilainen, Juha Heikkilä, and Arto Anttila, editors. 1995. *Constraint Grammar: a language-independent system for parsing unrestricted text*, volume 4 of *Natural Language Processing*. Mouton de Gruyter, Berlin and N. Y.
- Pasi Tapanainen and Timo Järvinen. 1997. A non-projective dependency parser. In *Proceedings of the 5th Conference on Applied Natural Language Processing*, Washington, D.C., April. ACL.
- Lucien Tesnière. 1959. *Éléments de syntaxe structurale*. Éditions Klincksieck, Paris.

A An example

The following sentence is excerpted from Jane Smiley's *Barn Blind* (HarperCollins, 1994, p. 8). The dependency links are enumerated and named, e.g. subj:>2 marks the subject of the head #2.

```

<ROOT> #0
<HE> @SUBJ PRON SG3 subj:>2
<WAS> @+FMAINV PAST #2 main:>0
<A> @DN> DET SG det:>4
<GIANT> @PCOMPL-S N SG #4 comp:>2
<WHO> @SUBJ PRON WH subj:>6
<PACED> @+FMAINV PAST #6 mod:>4
<AN> @DN> DET SG det:>9
<UNENDING> @A> A attr:>9
<CIRCLE> @OBJ N SG #9 obj:>6
<ABOUT> @<NOM PREP #10 mod:>9
<HIS> @A> PRON GEN SG3 attr:>12
<LITTLE> @QN> DET SG #12 qn:>13
<FARM> @<P N SG #13 pcomp:>10
<$,>
<VIEWING> @-FMAINV PCP1 #15 man:>6
<IT> @OBJ PRON ACC SG3 obj:>15
<FROM> @ADVL PREP #17 sou:>15
<EVERY> @QN> DET qn:>19
<ANGLE> @<P N SG #19 pcomp:>17
<IN> @<NOM PREP #20 mod:>19
<ALL> @QN> DET qn:>22
<WEATHERS> @<P N PL #22 pcomp:>20
<$,>
<AND> @CC CC cc:>15
<HE> @SUBJ PRON SG3 subj:>26
<WAS> @+FMAINV PAST #26 cc:>15
<A_LITTLE> @QN> DET SG qn:>28
<BOY> @PCOMPL-S N SG #28 comp:>26
<$,>
<WHO> @SUBJ PRON WH #30 subj:>43
<$,>
<WHENEVER> @ADVL ADV WH tmp:>35
<HE> @SUBJ PRON SG3 subj:>34
<DID> @+FAUXV PAST #34 v-ch:>35
<CATCH> @-FMAINV INF #35 mod:>30
<A> @DN> DET SG det:>37
<GLIMPSE> @OBJ N SG #37 obj:>35
<OUT_OF> @ADVL PREP #38 sou:>35
<THOSE> @DN> DET PL det:>41
<SIX-FOOT> @A> N SG attr:>41
<WINDOWS> @<P N PL #41 pcomp:>38
<$,>
<COULD> @+FAUXV AUXMOD #43 v-ch:>45
<HARDLY> @ADVL ADV >45
<BELIEVE> @-FMAINV INF #45 mod:>28
<WHAT> @OBJ PRON WH obj:>48
<HE> @SUBJ PRON SG3 subj:>48
<SAW> @+FMAINV PAST #48 obj:>45
<$,> punct:>48
  
```