AUTOMATIC F-STRUCTURE ANNOTATION FROM THE AP TREEBANK

Louisa Sadler
University of Essex
louisa@essex.ac.uk

Josef van Genabith
Dublin City University
josef@compapp.dcu.ie

Andy Way
Dublin City University
away@compap.dcu.ie

Proceedings of the LFG00 Conference

University of California, Berkeley

Miriam Butt and Tracy Holloway King (Editors)

2000
CSLI Publications
http://cslipublications.stanford.edu/

Abstract

We present a method for automatically annotating treebank resources with functional structures. The method defines systematic patterns of correspondence between partial PS configurations and functional structures. These are applied to PS rules extracted from treebanks. The set of techniques which we have developed constitute a methodology for corpus-guided grammar development. Despite the widespread belief that treebank representations are not very useful in grammar development, we show that systematic patterns of c-structure to f-structure correspondence can be simply and successfully stated over such rules. The method is partial in that it requires manual correction of the annotated grammar rules.

1 Introduction

The present paper discusses a method for the automatic annotation of treebanks with functional structures. A companion paper (Frank 2000) presents an alternative method for the automatic annotation of corpus resources. These closely related, but interestingly different methods have developed through much collaborative interchange. We present them in two separate contributions to allow for more in-depth discussion and comparison. We first describe our method and then exemplify its application to a grammar of 330 rules derived from a fragment of the AP treebank. We give some results concerning precision and recall for this grammar.

Treebanks which encode higher-level functional structure information in addition to phrase structure information, are required as training resources for probabilistic unification grammars and data-driven parsing approaches, e.g. (Bod and Kaplan 1998). Manual construction of such treebanks is very labour and cost intensive. As an alternative, one could envisage the construction of new, or the scaling-up of existing, unification grammars which could then be used to analyze corpora. However, these approaches are equally labour and cost intensive. What is more, even if a large-coverage unification grammar is available, typically, for each sentence it would come up with hundreds or thousands of candidate analyses from which a highly trained expert has to select. Although proposals have been made for filtering and ranking parsing ambiguities (e.g. (Frank et al. 1998)), to date none is guaranteed to uniquely determine the best analysis. In order not to compromise the quality of the corpus under construction, a linguistic expert is required to find the best among a large number of candidate analyses.

As a partial response to this data problem, van Genabith et al. (1999a,b,c) introduce a method for bootstrapping the construction of grammars from treebank resources. Their basic idea is the following: take an existing treebank, read off the CF-PSG following (Charniak 1996), manually annotate it with f-structure annotations, provide macros for the lexical entries and then "reparse" the treebanked trees simply following the original c-structure annotations. During this reparsing process, the f-structure annotations are resolved, and an f-structure is

produced. The process is deterministic if the annotations are, and to a large extent costly manual inspection of candidate analyses is avoided. The method successfully allows the creation of grammar resources but still involves one labour intensive manual component, namely annotation of the grammar rules with functional information. Much recent work in LFG, however, has shown that the c-structure f-structure correspondence for a configurational, generally endocentric language such as English, is largely predictable from a small set of mapping principles (King 1995, Kroeger 1995, Bresnan 2000). In the approach of Bresnan (2000) and colleagues, the mapping principles assume a highly articulated set of X'-schemata involving both functional and lexical projections in a configurational language such as English. A similar, but largely implicit, assumption about the predictability of the c- to f-structure mapping is also present in the earlier work in LFG (Kaplan and Bresnan 1982) where it turns up essentially as constraints on pairings of categories and grammatical functions (e.g. COMP is only appropriate for S, only NPs/DPs are OBJs and so forth). In general, the correspondence between c-structure and f-structure follows from linguistically determined principles which are partly universal, and partly language specific (Bresnan 2000), (Dalrymple 1999).

In the light of this, an obvious strategy to pursue is to implement a set of principles to automatically provide f-structure annotations of CFG rules derived from treebank representations, eliminating the manual step in the previous method. As a side effect, this can be expected to cast light on the soundness, accuracy and appropriacy of the linguists' generalisations: that is, the automatic procedure as applied to a large ruleset derived from a treebank, can serve as a potentially interesting testbed for the linguistic principles.

This paper substantially extends the research in van Genabith et al. (1999a,b,c) by showing how f-structure annotations of grammar rules extracted from treebanks may (to a large extent) be automated. The basic idea is very simple. We read off a CFG treebank grammar, using the first 100 trees of the AP treebank (Leech and Garside 1991). Systematic correspondences between elements in the c-structure domain and elements in the f-structure domain are then defined in general annotation templates. A corrected/completed version of this grammar is then used to induce f-structure assignments for PS trees from the treebank following the reparsing method of van Genabith et al. (1999a,b,c). The method is partial in that it requires manual inspection and correction of the output produced by the automatic annotation process. The method results in a set of annotated rules for real text.

The challenge for our for our approach is provided by the overtly flat analyses provided by the treebank input data, which in general do not conform to strongly hierarchical and recursive X' design principles. This has the effect that often what should be a single separate constituent is not assigned a corresponding *subtree* but merely a *substring* in the RHS of a flat treebank grammar rule. Our feature structure annotation principles underspecify rule RHSs and aim to pick out suitable substrings in flat rule RHSs.

Annotation principles express linguistic generalisations. The number of principles is substantially lower than the number of annotated CF-PSG grammar rules. The potential benefits of automation using annotation principles are considerable: substantial reduction in devel-

opment effort, hence savings in time and cost for treebank annotation and grammar development; the ability to tackle larger fragments in a shorter time, a considerable amount of flexibility for switching between different treebank annotation schemes, and a natural approach to robustness. The method we present may be viewed as a corpus-guided grammar development methodology.

The paper is structured as follows. In Section 2 we introduce the formalism for writing annotation templates. In Section 3 we discuss in some detail the NP fragment of our grammar and present a number of the templates involved. Section 4 presents the design of the automatic annotation experiment and evaluates the results obtained. Finally we conclude and outline further work.

2 Automatic f-structure annotation of CF Rules

Treebank grammars (CFGs extracted from treebanks) are very large and grow with the size of the treebank (Charniak 1996), (Krotov et al. 1998). They feature flat rules, many of which share and/or repeat significant portions of their RHSs. This causes several problems for manual annotation approaches such as the one described in van Genabith et al. (1999a,b,c). Annotation is labour intensive and repetitive, because of the sheer size and similarity of the rules, and annotation of rules on a one by one basis means that generalisations known to the annotator are simply not expressed. Of course, if the cardinality of the ruleset continues to grow with the size of the treebank, so too will the manual annotation task.

In LFG the correspondence between functional and constituent structure is partly defined in terms of annotations associated with c-structure nodes. Annotation follows universal and language specific principles. We can define principles as involving *partial* phrase structure configurations and apply them to all CFG rules that meet the relevant *partial* configuration. To give a simple example: a head principle assigns $\uparrow = \downarrow$ to the X daughter in all XP $\rightarrow \ldots$ X $\ldots$ configurations, irrespective of the surrounding categorial context. Such annotation principles capture generalisations, which can be used to *automatically* annotate PS configurations with functional structures in a highly general and economical way.

2.1 Feature Description Templates

In our approach to automatic annotation of c-structure rules, the linguist states generalisations over local sub-trees in the form of possibly partial and underspecified annotation principles, which take the following form:

```
Rhs > Lhs @ Anno
```

`Rhs > Lhs` is a possibly partial and underspecified CF-PSG grammar rule description using regular expressions, `Anno` is a set of feature structure annotations. To give a simple example, the following three annotation principles for `vp` rules state that the leftmost `v0` is the head of `vp`, that an `np` following a `v0` is a direct object and that a sequence of two `v0`s (as in the flat treebank analyses of auxiliary/modal constructions) induces open complement `xcomp` feature structures where the subject of the complement is controlled by the subject of the superordinate feature structure.

```
vp:VP > v0:V *
      @ [VP === V]
vp    > * v0:V0 np:NP *
      @ [V0:obj === NP]
vp    > * v0:V0 v0:V1 *
      @ [V0:xcomp === V1, V0:subj === V0:xcomp:subj]
```

Given a flat (treebank) CF-PSG rule (for strings like *did order a recount*) of the form

```
vp:VP > v0:V0 v0:V1 np:NP
```

these annotation principles conspire to induce the following feature structure annotation, as required:

```
vp:VP > v0:V0 v0:V1 np:NP
      @ [VP === V0, V1:obj === NP,
         V0:xcomp === V1, V0:subj === V0:xcomp:subj]
```

Our CF-PSG rule description language consists of regular expressions including Kleene star `*`, positive Kleene `+`, optionality `?`, disjunction `|` and a limited form of complement `~`. Terminal symbols are either `category:feature structure` pairs (as in `np:NP`) or, for convenience, simple categories if the feature structure associated with the category is not mentioned in the feature structure annotations. The `*` and `+` operators can be used with or without arguments. Without arguments `*` denotes any string (including the empty string) while `+` denotes any string of lenght greater one.

Both `Lhs` and `Rhs` in annotation principles are regular expressions. The interpretation of an annotation principle such as

```
vp    > * v0:V0 v0:V1 *
      @ [V0:xcomp === V1, V0:subj === V0:xcomp:subj]
```

is: if you find a sequence of two adajecent `v0`s in the RHS of a `vp` rule then annotate that rule with the feature structure annotations stated in the `@ Anno` part of the annotation principle. A single CF-PSG rule may receive annotations from more than one (partial) annotation

principle. What is more, a single annotation principle may match a single CF-PSG rule more than once. Consider the following flat `vp` rule:

```
vp:VP > v0:V0 v0:V1 v0:V2 np:NP
```

for a string such as *may have ordered a recount*. The rule RHS features two adajecent `v0` sequences and the `v0` sequence annotation template is going to match twice. The feature structures induced by all such matches are collected and the rule is annotated accordingly:

```
vp:VP > v0:V0 v0:V1 v0:V2 np:NP
      @ [VP === V0, V1:obj === NP,
         V0:xcomp === V1, V0:subj === V0:xcomp:subj,
         V1:xcomp === V2, V1:subj === V1:xcomp:subj]
```

As a final example, consider a flat (treebank) CF-PSG rule for a string such as *may have ordered the county to recount the ballot*.

```
vp:VP > v0:V0 v0:V1 v0:V2 np:NP infp:I
```

The final infinitival phrase *to recount the ballot* is an open complement argument to the rightmost `v0`. In our rule description language this can be expressed as follows:

```
vp > * v0:V0 *(~v0) infp:I *
     @ [V0:xcomp === I]
```

This annotation principle matches `vp` rules containing an `infp` constituent. This constituent provides an `xcomp` to a `v0` such that no other `v0` (though possibly other constituents) may intervene between it and the `infp`:

```
vp:VP > v0:V0 v0:V1 v0:V2 np:NP infp:I
      @ [VP === V0, V1:obj === NP,
         V0:xcomp === V1, V0:subj === V0:xcomp:subj,
         V1:xcomp === V2, V1:subj === V1:xcomp:subj,
         V2:xcomp === I]
```

Notice that the principle does not specify a subject function for the open complement. This is provided lexically by a feature structure macro for object control verbs such as *order*.

Our annotation principles factor out and express generalisations over the flat treebank CF-PSG rules. The generalisations are expressed in terms of partial and underspecified rule descriptions. The annotation principle compiler applies principles to the CF-PSG rules extracted from the treebank. Statement and processing of the annotation principles is order independent. The compilation of principles over grammar rules is efficient. Our current implementation annotates about 50 CF-PSG rules per second.

3 The NP Grammar and Templates

In the previous section, we introduced the formalism for writing annotation templates. To give a flavour of what is involved, in this section we will present several aspects of the NP grammar. We show that this approach permits the linguist to state simple generalisations and translate them straightforwardly into templates. In our work to date we have developed templates for the entire grammar of 330 rules derived from the treebank. However, we have chosen to concentrate on one section of the grammar for expository purposes.

The NP fragment constitutes the largest and most complex set of phrase structure rules induced for a single non-terminal category from our set of sentences. Because of its size and complexity, and because the issues which it raises give a good feel for what is involved in our approach to automatic annotation, we limit discussion to this fragment. The grammar fragment contains 142 rules, for which we have written 29 templates. 23 categories are attested within NP, a very high proportion of the overall number of categories in the grammar, which is 41 (29 lexical and 12 non-terminal categories)[1].

(1) Categories found within NP

`det`	`ndet`	`adj`	`adjp`	`dadj`
`n0`	`np`	`num`	`title`	`posspron`
`pron`	`pnct`	`conj`	`relcl`	`pp`
`p`	`ntadv`	`adv`	`v0`	`vp`
`fn`	`tgp`	`infp`		

3.1 Compaction and Supercategories

In earlier work (van Genabith et al. 1999c) we found that the rich set of tags used in the AP treebank provided much useful f-structure information which could be simply re-expressed in a set of lexical macros. The distinctions introduced by the AP tagset, in common with other tagsets, are extremely fine-grained because all sorts of subcategorial distinctions are expressed by means of the monadic category labels. In very many cases, these subcategorial distinctions are ones which would be expressed by means of grammatical features at f-structure in LFG (distinctions such as number, verbform, and so on). Since this information is recaptured by means of these lexical macros, they hypothesized that it would be helpful to abstract away from the specificities of the particular set of tags used in their database of sentences in favour of a smaller set of "supertags" in order to develop a stand-alone resource. This can be viewed as plugging holes in the grammar, for it permits a more general grammar to be derived from that which would otherwise be read off from the treebank entries. Therefore van Genabith et al. (1999b) introduce a structure-preserving grammar compaction method which first uses lexical macros to associate f-structure constraints with the words in

[1]The category set that we are dealing with is derived from the original AP tagset by a process of compaction, which we describe in the following section.

the tree and then, having specified a mapping between tags and "supertags" (generalisations over tags), uses the latter to reparse the treebank entries and compile a "generalised" CFG from the tree using the method of (Charniak 1996).

The work described here investigates the automatic association of f-structure constraints with the rules of the CFG by means of annotation templates. We have found that the categorial compaction described in (van Genabith et al. 1999b) has provided an excellent basis for automatic f-annotation: many of the distinctions preserved in the reduced (generalised) tagset are precisely those which we require to guide automatic annotation. For example, in the nominal domain, the large number of distinctions made between nominal elements on the basis of morphosyntactic class membership are eliminated, but the distinction of nouns with adverbial function is maintained. We make use of such information to directly guide the f-structure annotation process. Likewise, the AP tagset makes a series of distinctions within the verbal/sentential system which, suitably generalised over, are useful in the same way. As an example, we assign supertags over sets of AP tags as indicated below:

(2)

Supertag	AP Tag	Description
FA	Fa	Adverbial clause
	Fa&	First conjunct of an adverbial clause
	Fa+	Second conjunct of an adverbial clause
FN	Fn	Noun clause
	Fn&	First conjunct of a noun clause
	Fn+	Second conjunct of a noun clause
RELCL	Fr	Relative clause
	Fr&	First conjunct of a relative clause
	Fr+	Second conjunct of a relative clause
INFP	Ti	to + infinitive clause
	Ti&	First conjunct of a to + infinitive clause
	Ti+	Second conjunct of a to + infinitive clause

The following, exceptionless generalisations can be stated about these derived categories.

(3) An FN within NP is a COMP in the NP's f-structure

```
np:NP > * fn:FN *
      @ [NP:comp === FN]
```

(4) An infinitival VP within NP is an XCOMP in the NP's f-structure

```
np:NP > * infp:I *
      @ [NP:xcomp === I]
```

(5) A RELCL within NP is a RELMOD in the NP's f-structure

```
np:NP > * relcl:R *
      @ [NP:relmod === R]
```

Of course, not all constituents which map to the f-structure function RELMOD *are* represented as **relcl** in the treebank entries. The sample of 100 sentences contains a number of cases of reduced relative clauses, which are associated with the (super-)category **vp** in our collapsed tagset. Given the distinctions made in the verbal supertag set, the following generalisation may be made about the occurrence of the (super-)tag **vp** within the noun phrase:

(6)
```
np:NP > * vp:R *
      @ [NP:relmod === R]
```

Notice that the two **relcl** annotation principles can be collapsed into a single principle as follows:

(7)
```
np:NP > * (relcl:R|vp:R) *
      @ [NP:relmod === R]
```

Since the AP tagset encodes adverbial function, the supertag **ntadvp** (for nominal temporal adverbial) can be straightforwardly related to a specific function:

(8) An NTADVP maps to an NP_ADJUNCT in the mother's f-structure

```
np:NP > * ntadvp:NT *
      @ [NP:np_adjunct:el === NT]
```

The original AP tagset contains more than 20 pronominal tags, which we collapse to two supertags: **posspron** for possessive pronouns, and **pron** for all other pronouns. Again, the c-structure to f-structure mapping templates are simple to write for these categories[2]:

(9)
```
np:NP > * posspron:P *
      @ [NP:poss === P]
```

(10)
```
np:NP > * pron:P *
      @ [NP === P]
```

To take a more complicated example, the AP tagset distinguishes the following subtypes of Adjectives: **ja jb jj da da2 dar dat**. All the "j" tags are adjectives, either predicative, central and attributive. The "d" adjectives are "after determiners" such as "former, such, few, several...". The "j" adjectives are attributive modifiers within NP, and under our treatment,

[2]Of course, given the flatness of the trees, a **posspron** might denote a POSS function, but not necessarily the POSS of the f-structure of its mother node (it might be more deeply embedded in the f-structure). However, in our fragment this is not attested and we can make do with the simple generalisation in (9).

correspond to NP_ADJUNCT and HEADMOD grammatical functions,[3] while the "d" adjectives map to SPEC or may serve as the head of NP in the absence of a nominal element. In our collapsed tagset, all the "d" adjectives are treated as **dadj** and all the "j" adjectives as **adj**: this distinction, which is a simple generalisation of the categorial distinctions made in the treebank tags, corresponds to a difference in grammatical functional possibilities for these subtypes of adjectives. For **dadj**, the generalisation that we wish to state is that it maps to SPEC if there is a nominal f-head, otherwise to f-head.

(11)
```
np:NP > * dadj:DA * (n0|np|num) *
        @ [NP:spec === DA]
```

(12)
```
np:NP > * dadj:DA *(~(n0|np|num))
        @ [NP === DA]
```

This pair of templates is essentially equivalent to annotating a **dadj** node with a disjunction $(\downarrow \in (\uparrow \mathrm{ADJ})) \vee (\downarrow = \uparrow)$. We recognise the ability of these adjectives to stand on their own as the head of NP by permitting the category **dadj** to serve as the head (when no other potential head is present) rather than by reassigning them to a nominal or determiner category.

The other subclass of adjectives serve as nominal modifiers within NP. The LFG treatment of attributive adjectives is as members of the set-valued feature ADJUNCT (here NP_ADJUNCT). This is appropriate for iterative uses of adjectives which separately restrict the interpretation of the head noun. However, our corpus contains a significant number of cases in which an adjective may appear on the left periphery (and part) of what is essentially a complex (internally-modified) nominal head, as in *jump shot, national guard troops, wide area telephone service*. For the latter cases we have used the additional grammatical function HEADMOD: the prototypical use of this function is in cases of noun-noun compounding, which abound in our small corpus extract. We shall have more to say about the HEADMOD function when we discuss NN compounds below.

Treating adjectives as potentially mapping to HEADMOD as well as to NP_ADJUNCT leads to the following template information for **adj**: adjectives next to nominal heads are either NP_ADJUNCTs or HEADMODs, other adjectives are NP_ADJUNCTs.

(13)
```
np:NP > * adj:A (n0:N|np:N|num:N) *
        @ [ ( NP:np_adjunct:el === A ;
              N:headmod === A ) ]
```

(14)
```
np:NP > * adj:A ~(n0|np|num) *
        @ [NP:np_adjunct:el === A]
```

Since only single, uncomplemented and unmodified adjectives can be used as part of these sorts of structures, the template for AP is simple to state: it maps to the nominal adjunct function.

[3] We discuss this distinction at greater length below.

```
(15)  np:NP > * adjp:A *
            @ [NP:np_adjunct:el === A]
```

3.2 PP Dependents

Distinguishing OBL from ADJUNCTs is a notoriously difficult problem, especially within NPs, where the PP dependents are largely optional. One possible tack would be to treat all PP dependents of nominals as ADJUNCTs (this approach is adopted in (Butt et al. 1999)), that is, to treat all optional arguments of nominal heads as syntactic modifiers (some of which will be semantic arguments) rather than syntactic arguments. On this view, our annotation templates would simply need to state the generalisation that the category `pp` maps to the ADJUNCT function. However, inspection of our set of sentences suggests that while the second of two PPs is always an ADJUNCT, a PP adjacent to the nominal head may be either an OBL or an ADJUNCT. Although it is claimed that OBLIQUE and ADJUNCT PPs can reorder rather freely, the strings in the template reflect the ordering which would be imposed by the X'-schemata. The following templates, therefore, introduce a measure of disjunction into the annotation process:

```
(16)  np:NP > * pp pp:P *
            @ [NP:np_adjunct:el === P]
```

```
(17)  np:NP > * (no|np|num) pp:P *
            @ [ ( NP:np_adjunct:el === P ;
                  NP:obl === P ) ]
```

3.3 Head Modifier Structures

The treebank representations of NPs are very flat and often quite complex - the following are representative.

```
(18)   np:A > [det:B,adj:C,adj:D,n0:E,n0:F,adjp:G]
       np:A > [det:B,adj:C,n0:D,n0:E,relcl:F]
       np:A > [det:B,adj:C,n0:D,n0:E]
       np:A > [det:B,adj:C,n0:D,ntadvp:E,relcl:F]
       np:A > [det:B,adj:C,n0:D,pnct:E,vp:F]
       np:A > [det:B,adj:C,n0:D,pp:E]
       np:A > [det:B,dadj:C,n0:D,n0:E,n0:F,relcl:G]
       np:A > [n0:B,n0:C,n0:D,n0:E,vp:F]
       np:A > [n0:B,n0:C,n0:D,n0:E]
       np:A > [n0:B,n0:C,n0:D,np:E]
       np:A > [n0:B,n0:C,n0:D,ntadv:E,pp:F)]
```

```
np:A > [n0:B,n0:C,n0:D]
np:A > [n0:B,n0:C,np:D,pnct:E,np:F]
np:A > [n0:B,n0:C,np:D]
np:A > [n0:B,n0:C,pnct:D,np:E]
np:A > [n0:B,n0:C,pnct:D,relcl:E]
```

Given the remarkable paucity of internal structure here, a major issue is determining what category is the head of NP; that is, what category is to be annotated (↑= ↓A striking feature of many of the NP rules is that they contain strings of nominal categories. In such cases, these flat strings of nominal categories behave in an essentially right-headed fashion. The elements **n0**, **num** and **np** typically serve as the head and the rightmost such element present (provided it is not preceded by **pnct**, which marks an appositional structure), is the head of the NP. This generalisation can be stated as follows:

(19)
```
np:NP > *(~conj) ~(conj|pnct) (n0:N|np:N|num:N)
          *(~(n0|np|num))
        @ [NP === N]
```

As can be seen, this template must take into account the complicating factor of coordination, and in particular, the flat representation of coordination, which we discuss in the following section. The fragment of grammar in (18) and the template (19) illustrate nicely a feature of treebank representations, namely their extremely flat representations. The template above must search for the rightmost category appropriate to serve as head, from a sequence of categories.

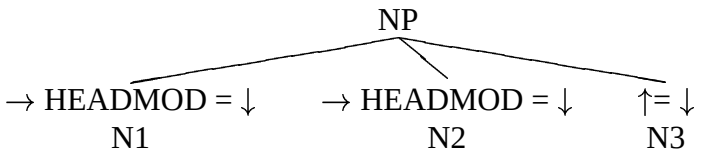
The overly flat treebank representations are very problematic when it comes to determining the correct head-modifier relationships within the noun phrase. One possibility is to treat all pre-head nominal (and adjectival) elements as direct modifiers of the final nominal head. It would be trivial to then define the appropriate annotation template mapping all such prehead nominal modifiers into a set-valued ADJUNCT f-structure. This would entail, for example, treating *law enforcement officer* as a head *officer* modified by a set of adjuncts { *law, enforcement* }. This is essentially the approach adopted in the LFG grammars of the PARGRAM project described in (Butt et al. 1999). The difficulty with this is that a flat representation as ADJUNCTS at f-structure would fail to encode the semantic modification relations which hold within these pre-head modifiers (although, of course, it is possible to keep trace of at least linear position in the string of modifiers by judicious indexing of the elements in the ADJUNCTS set).

These structures are extremely common in our fragment: for example, there are 52 rules containing a total of 72 simple **n0 n0** sequences in which nominal elements modify nominal structures to their right. We treat these as head-modifer structures, introducing a new (single-valued) grammatical function HEADMOD. Strings such as *guard helicopters*, *smoke inhalation*, *hospital spokesman*, *law enforcement officers*, and many others in our sample, may be viewed as syntactic structures (each word is associated with a separate terminal category in the treebank representations) which are built according to "morphological" principles

(number is marked on the final element, the structures are head final), in which each element modifies the f-structure of the element to its right.

This approach is also problematic, however. A nominal is taken to modify the f-structure projected from its nominal sister, as shown schematically in (20), where (→) denotes the f-structure of the immediately adjacent right sister, with the corresponding schematic f-structures in (21).

(20)



(21)

$$\begin{bmatrix} \text{PRED} & \text{`N3'} \\ \text{HEADMOD} & \begin{bmatrix} \text{PRED} & \text{`N2'} \\ \text{HEADMOD} & \begin{bmatrix} \text{PRED} & \text{`N1'} \end{bmatrix} \end{bmatrix} \end{bmatrix}$$

It is clear, however, that such "left-branching" structures are not always correct for the strings in our corpus, and that in some cases a "right-branching" f-structure, with a complex head (as shown in (22) would be more correct[4]. These sorts of structures, with what amounts to complex PREDs, are not permissible in LFG.

(22)

$$\begin{bmatrix} \text{HEAD} & \begin{bmatrix} \text{PRED} & \text{`N3'} \\ \text{HEADMOD} & \begin{bmatrix} \text{PRED} & \text{`N2'} \end{bmatrix} \end{bmatrix} \\ \text{HEADMOD} & \begin{bmatrix} \text{PRED} & \text{`N1'} \end{bmatrix} \end{bmatrix}$$

For the moment, however, the template simply picks out sequences of **n0** categories and adds the f-structure constraint that the first is the HEADMOD of the second. Note that the template cannot, of course, equate the f-structure of the mother with the rightmost category in the pair, since the f-structure of this category may itself be a HEADMOD within a containing f-structure: picking out the head of NP is performed by the head template given in (19) above.

(23)
```
np > * n0:N1 n0:N0 *
     @ [N0:headmod === N1]
```

[4]Nevertheless, it must be stated that such NPs in our grammar are all treated as 'left-branching' structures. As we point out, this means that in some cases we provide the wrong treatment for such phenomena. The results given later in the paper were performed on this 'faulty' set of NPs. We are confident that reinterpreting these N-N compounds correctly will not prevent us from achieving equally good figures for precision and recall, but we have yet to rewrite the grammars and templates as desired, so that this must remain as speculation at this stage.

What else, apart from **n0**, maps to the headmod function? Members of the category **num**, like **adj** may be either ADJUNCTs or HEADMOD: the template is shown in (24). We extend the same treatment to titles, as in (25).

(24)
```
np > * num:N n0:N0 *
   @ [ (N0:headmod === N ; N0:np_adjunct:1 === N) ]
```

(25)
```
np > * title:T n0:N0 *
   @ [N0:headmod === T]
```

3.4 Coordination and Flat Trees

The approach to constituent coordination in LFG treats the conjuncts as a set at f-structure, with the conjunction contributing a value directly to the semantic structure. Our formalism does not currently support set values, and we model constituent coordination by treating the conjunction as a predicate taking a CONJ argument which itself takes any number of indexed arguments. The treatment of the conjuncts themselves then closely resembles our treatment of the set valued feature ADJUNCT, as (26) illustrates.

(26)
```
np:A > [np:B,pnct:C,np:D,pnct:E,conj:F,np:G]
```

$$\begin{bmatrix} \text{PRED} & \text{and} \\ \text{CONJ} & \begin{bmatrix} 1 & [\text{PRED} \quad \ldots.] \\ 2 & [\text{PRED} \quad \ldots.] \\ 3 & [\text{PRED} \quad \ldots.] \end{bmatrix} \end{bmatrix}$$

Ideally, then, the template must pick out the **conj** as the f-head and treat other categories, except for **pnct** as CONJ functions. However, consideration of the set of np rules involving coordination makes clear that things are unfortunately considerably more complicated. Because the treebank representations are extremely flat, the scope of coordination is not indicated by the presence of a distinct sub-tree: this means that it is not possible to assign all daughters (except **conj** and **pnct** to CONJ functions.

(27)
```
np:A > [n0:B,conj:C,n0:D)]
np:A > [np:B,conj:C,np:D)]
np:A > [np:B,pnct:C,conj:D,np:E)]
np:A > [np:B,pnct:C,np:D,conj:E,np:F)]
np:A > [np:B,pnct:C,np:D,pnct:E,conj:F,np:G)]
np:A > [adj:B,conj:C,adj:D,n0:E,n0:F)]
```

```
np:A > [adj:B,conj:C,adj:D,n0:E,pp:F)]
np:A > [adv:B,num:C,adj:D,n0:E,n0:F,conj:G,n0:H)]
np:A > [det:B,adj:C,n0:D,conj:E,n0:F,n0:G)]
np:A > [det:B,adj:C,n0:D,conj:F,n0:G,pp:E)]
np:A > [det:B,adv:C,conj:D,adv:E,adj:F,n0:G,pnct:H,
        relcl:I)]
np:A > [det:B,n0:C,n0:D,pnct:E,n0:F,conj:G,n0:H)]
np:A > [posspron:B,n0:C,pp:D,conj:E,adv:F)]
```

The **n0** conjunction principle assigns `n0`s preceding a `conj n0` sequence as individual conjuncts. It matches as many times as there are `n0`s preceding the `conj n0` sequence, the resulting annotations are collected and the rule is annotated accordingly. Notice that this can result in multiple but identical `C:conj:el === N0, NP === C` annotations which does not cause any harm. However, because of the flat treebank rules the **n0** conjunction principle is no more than an approximation. It can introduce errors in case two adjacent `n0`s should be analysed as `headmod` structures rather than as coordinate elements in a coordinate structure.

(28)
```
np:NP > * n0:Nx * conj:C n0:N0 *
        @ [C:conj:el === Nx, C:conj:el === N0,
           NP === C]
```

Further complications are the coordination of adjectives directly under np and even of adverbs modifying adjectives under np. General statements can be written for these under which in each case the coordinate structure is identified and assigned the correct sort of ADJUNCT function in the mother f-structure.

(29)
```
np:NP > * adj:Ax * conj:C adj:A *
        @ [C:conj:el === Ax, C:conj:el === A,
           NP:np_adjunct:el === C]
```

4 Experiments

In this section, we report on experiments in automatically compiling the templates over the grammar rules and compare the results to our hand-coded grammar.

4.1 Experiment Design and Data

Our experiment involves the first 100 trees of the AP treebank (Leech and Garside 1991). We preprocess the treebank using the structure preserving grammar compaction method reported in (van Genabith et al. 1999b) and extract a treebank grammar following (Charniak 1996).

The large number of highly discriminating terminal and non-terminal categories results in a large number of often very specific rules: the grammar compaction method provides a more general grammar that still preserves all important categorial information to drive automatic annotation. Compaction works by generalising tags, i.e. collapsing tags (and categories) into supertags. This reduces the number of rules from 509 to 330. The sentences in the fragment range from 4 to 50 terminal tokens (including punctuation symbols). We develop a set of feature structure annotation templates. Our template interpreter compiles the templates over the rules in the treebank grammar.

In order to evaluate the results of automatic annotation we manually constructed a reference grammar following (van Genabith et al., 1999a,b,c). The grammar features 1128 annotations, on average about 3.4 annotations per rule.[5]

4.2 Automatic Annotation and Evaluation

We constructed 129 templates, this against 330 CFG rules resulting in a template/rule ratio of 0.39. We expect the ratio to skew in favour of templates as we proceed to larger fragments. Automatic annotation generates 1108 annotations, on average about 3.36 annotations per rule. We evaluate the automatic annotation procedure in terms of *precision* and *recall*.

(30)

$$precision = \frac{\#generated\ annotations\ also\ in\ reference}{\#generated\ annotations}$$

$$recall = \frac{\#reference\ annotations\ also\ generated}{\#reference\ annotations}$$

	Experiment
precision	93.38%
recall	91.58%

These numbers, although good, are conservative: *precision* and *recall* are computed automatically and currently our annotation matcher is not complete.[6]

The results are encouraging and indicate that while automatic annotation is (slightly) more often partial than incorrect, a small number of annotation templates can be written for a

[5] Templates, grammars and f-structures are available at `http://www.compapp.dcu.ie/~away/Treebank/treebank.html`.

[6] E.g.: `P1 = P2` $\models$ `P2 = P1` and `A=B, A:P1 = P2` $\models$ `B:P1 = P2` where `Pi` are paths and `A,B` variables; currently our precision and recall programme misses `P1:P2 = P3, P1 = A` $\models$ `A:P2 = P3` type inferences.

grammar fragment which appears to be quite complex. The generalisations made are simple and robust, and can be expected to considerably ease the annotation burden on the grammar writer.

Having been confronted with 'real' text, we have been forced to distinguish a number of grammatical functions for which there is good c-structure evidence and/or motivation in real text, but which are not discussed in the theoretical literature. In particular, we have postulated a pre-modificational HEADMOD grammatical function within NPs. The biggest challenge presented by the very flat treebank representations concerns the coordination data, and in particular the interaction of coordination with other phenomena.

5 Conclusions and Further Work

We have presented and extensively exemplified a method for the automatic f-structure annotation of treebank grammars. At this stage our intent has been to present the methods and to explore some of their potential. The approach applies to a CFG, such as that derived from a treebank, and yields an annotated grammar, which can either be used to reparse treebank trees to induce feature structure annotations for treebank trees or serve as a basis for developing a stand-alone LFG resource. It uses a compaction technique for generalising overspecific categorisation. The structure of treebank entries remains unchanged. We implemented an order-independent annotation template interpreter. Order independence can ease development and maintainance of annotation principles, but requires more complex rule constraints.

Automatic annotation holds considerable potential in curtailing development costs and opens up the possibility of tackling large fragments. To date, our experiments are admittedly small-scale. Still, we have presented an important grammar development and treebank annotation methodology which is data-driven, semi-automatic and reuses existing resources. We found the LFG framework very conducive to our experiments. We do believe, however, that the methods can be generalised, and we intend to apply them in an HPSG scenario. Note further that our methods encourage work in the best linguistic tradition as (i) they are concerned with real language and (ii) they enforce generalisations in the form of annotation principles. The experiments show how theoretical work and ideas on principles can translate into grammar development for real texts. In this sense the methods bridge the often perceived gap between theoretically motivated views of grammar as a set of principles versus grammars for 'real' text.

Bibliography

Bod, R., and R. Kaplan. 1998. A Probabilistic Corpus-Driven Model for Lexical-Functional Analysis. In *Proceedings of COLING/ACL'98*, 145–151.

Bresnan, J. 2000. Lexical Functional Syntax. Forthcoming, Blackwells Publishers, Oxford.

Butt, M., T. H. King, M.-E. Nino, and F. Segond. 1999. *The Grammar Writer's Cookbook*. Stanford: CSLI Publications.

Charniak, E. 1996. Tree-bank Grammars. In *AAAI-96. Proceedings of the Thirteenth National Conference on Artificial Intelligence*, 1031–1036. MIT Press.

Dalrymple, M. 1999. Lexical-Functional Grammar. Manuscript, Xerox PARC.

Frank, A., T. King, J. Kuhn, and J. Maxwell-III. 1998. Optimality Theory Style Constraint Ranking in Large-scale LFG Grammars. In M. Butt and T. King (Eds.), *Proceedings of the LFG98 Conference*, University of Queensland, Brisbane, CSLI Online Publications, Stanford, CA. `http://www-csli.stanford.edu/publications/.`

Frank, A. 2000. Automatic F-structure Annotation of Treebank Trees. In M. Butt and T. King (Eds.), *Proceedings of the LFG00 Conference*, CSLI Online Publications, Stanford, CA. `http://www-csli.stanford.edu/publications/.`

Kaplan, R., and J. Bresnan. 1982. Lexical Functional Grammar: a Formal System for Grammatical Representation. In J. Bresnan (Ed.), *The Mental Representation of Grammatical Relations*, 173–282. Cambridge, Mass: MIT Press.

King, T. H. 1995. *Configuring Topic and Focus in Russian*. Stanford: CSLI Publications.

Kroeger, P. 1995. *Phrase Structure and Grammatical Relations in Tagalog*. Stanford: CSLI.

Krotov, A., M. Hepple, R. Gaizauskas, and Y. Wilks. 1998. Compacting the Penn Treebank Grammar. In *Proceedings of COLING/ACL'98*, 699–703.

Leech, G., and R. Garside. 1991. *Running a Grammar Factory: On the Compilation of Parsed Corpora, or 'Treebanks'*. Mouton de Gruyter, Berlin. 15–32.

van Genabith, J., L. Sadler, and A. Way. 1999a. Data-Driven Compilation of LFG Semantic Forms. In *EACL'99 Workshop on Linguistically Interpreted Corpora (LINC-99), Bergen, Norway, June 12th*, 69–76.

van Genabith, J., L. Sadler, and A. Way. 1999b. Structure Preserving CF-PSG Compaction, LFG and Treebanks. In *Proceedings ATALA Workshop - Treebanks*, Journées ATALA, Corpus annotés pour la syntaxe, Universite Paris 7, France, 18-19 Juin 1999, 107–114.

van Genabith, J., A. Way, and L. Sadler. 1999c. Semi-Automatic Generation of F-Structures from Tree Banks. In M. Butt and T. King (Eds.), *Proceedings of the LFG99 Conference*, Manchester University, 19-21 July, CSLI Online Publications, Stanford, CA. `http://www-csli.stanford.edu/publications/.`