The design of a computer language for linguistic information

The Harvard community has made this article openly available. Please share how this access benefits you. Your story matters

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Published Version</td>
<td><a href="http://www.aclweb.org/anthology-new/P/P84/P84-1075.pdf;http://dx.doi.org/10.3115/980491.980566">http://www.aclweb.org/anthology-new/P/P84/P84-1075.pdf;http://dx.doi.org/10.3115/980491.980566</a></td>
</tr>
<tr>
<td>Citable link</td>
<td><a href="http://nrs.harvard.edu/urn-3:HUL.InstRepos:2309659">http://nrs.harvard.edu/urn-3:HUL.InstRepos:2309659</a></td>
</tr>
<tr>
<td>Terms of Use</td>
<td>This article was downloaded from Harvard University’s DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at <a href="http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA">http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA</a></td>
</tr>
</tbody>
</table>


The Design of a Computer Language for Linguistic Information
Stuart M. Shieber
Artificial Intelligence Center
SRI International
and
Center for the Study of Language and Information
Stanford University

Abstract

A considerable body of accumulated knowledge about the design of languages for communicating information to computers has been derived from the subfields of programming language design and semantics. It has been the goal of the PATR group at SRI to utilize a relevant portion of this knowledge in implementing tools to facilitate communication of linguistic information to computers. The PATR-II formalism is our current computer language for encoding linguistic information. This paper, a brief overview of that formalism, attempts to explicate our design decisions in terms of a set of properties that effective computer languages should incorporate.

1. Introduction

The goal of natural-language processing research can be stated quite simply: to endow computers with human language capability. The pursuit of this objective, however, has been a difficult task for at least two reasons: first, this capability is far from being a well-understood phenomenon; second, the tools for teaching computers what we do know about human language are still very primitive. The solution of these problems lies within the respective domains of linguistics and computer science.

Similar problems have arisen previously in computer science. Whenever a new computer application area emerges, there follow new modes of communication with computers that are geared towards such areas. Computer languages are a direct result of this need for effective communication with computers. A considerable body of accumulated knowledge about the design of languages for communicating information to computers has been derived from the subfields of programming language design and semantics.

2. The Critical Properties of the Language

It is not the purpose of this paper to provide a comprehensive description of the PATR-II project, or even of the formalism itself. Rather, we will discuss briefly the critical
properties of PATR-II to give a flavor for our approach to the design of the language. References to papers with more complete descriptions of particular aspects of the project are provided when appropriate.

2.1. Simplicity: An Introduction to the PATR-II Formalism

Building on a convergence of ideas from the linguistics and AI communities, PATR-II takes as its primitive operation an extended pattern-matching technique, unification, first used in logic and theorem-proving research and lately finding its way into research in linguistics [Kay, 79; Gazdar and Pullum, 82] and knowledge representation [Reynolds, 70; Ait-Kaci, 83]. Instead of unifying logic terms, however, PATR unification operates on directed acyclic graphs (DAG).3

DAGs can be atomic symbols or sets of label/value pairs, where the values are themselves DAGs (either atomic or complex). Two labels can have the same value—thus the use of the term graph rather than tree. DAGs are notated either by drawing the graph structure itself, with the labels marking the arcs, or, as in this paper, by notating the sets of label/value pairs in square brackets, with the labels separated from their values by a colon, e.g., a DAG associated with the verb "knight" (as in "Uther wants to knight Arthur") would appear (in at least one of our grammars) as

\[
\text{[cat: v}
\begin{align*}
\text{head: [aux: false} & \\
\text{form: nonfinite} & \\
\text{voice: active} & \\
\text{trans: [pred: knight} & \\
\text{arg1: <fl134>} & \\
\text{arg2: <fl138>} & \\
\text{syncat: [first: [cat: np} & \\
\text{head: [trans: <fl134>]} & \\
\text{rest: [first: [cat: np} & \\
\text{head: [trans: <fl138>]} & \\
\text{rest: <f1140> lambdas]}} & \\
\text{tail: <f1140>]} &
\end{align*}
\]

Reentrant structure is notated by labeling the DAG with an arbitrary label (in angle brackets), then using that label for future references to the DAG.

Associated with each entry in the lexicon is a set of DAGs.4 The root of each DAG will have an arc labeled cat whose value will be the category of the associated lexical entry. Other arcs may encode information about the syntactic features, translation, or syntactic subcategorization of the entry. But only the label cat has any special significance; it provides the link between context-free phrase structure rules and the DAGs, as explicated below.

PATR-II grammars consist of rules with a context-free phrase structure portion and a set of unifications on the DAGs associated with the constituents that participate in the application of the rule. The grammar rules describe how constituents can be built up to form new constituents with associated DAGs. The left side of the rule lists the cat values of the DAGs associated with the filial constituents; the right side, the cat of the parent. The associated unifications specify equivalences that must exist among the various DAGs and sub-DAGs of the parent and children. Thus, the formalism uses only one representation—DAGs—for lexical, syntactic, and semantic information, and one operation—unification—on this representation.

By way of example, we present a trivial grammar for a fragment of English with a lexicon associating words with DAGs.

\[
S \rightarrow \text{NP VP} \\
<\text{VP agr}> = <\text{NP agr}> \\
<\text{VP agr}> = <\text{v agr}> \\
\text{Uther:} \\
<\text{cat}> = \text{np} \\
<\text{agr number}> = \text{singular} \\
<\text{agr person}> = \text{third} \\
\text{Arthur:} \\
<\text{cat}> = \text{np} \\
<\text{agr number}> = \text{singular} \\
<\text{agr person}> = \text{third} \\
\text{knights:} \\
<\text{cat}> = \text{v} \\
<\text{agr number}> = \text{singular} \\
<\text{agr person}> = \text{third}
\]

This grammar (plus lexicon) admits the two sentences "Uther knights Arthur" and "Arthur knights Uther." The phrase structure associated with the first of these is:

\[
[\text{s [np Uther] [vp [v knights] [vp Arthur]]}]
\]

The VP rule requires that the agr feature of the DAG associated with the VP be the same as (unified with) the agr of the V. Thus, the VP's agr feature will have as its value the same node as the V's agr, and hence the same values for the person and number features. Similarly, by virtue of the unification associated with the S rule, the NP will have the same agr value as the VP and, consequently, the V. We have thus encoded a form of subject-verb agreement.

Note that the process of unification is order-independent. For instance, we would get the same effect regardless of whether the unifications at the top of the parse tree were effected before or after those at the bottom. In either case, the DAG associated with, e.g., the VP node would be
These trivial examples of grammars and lexicons offer but a glimpse of the techniques used in writing PATR-II
grammars, and do not begin to employ the power of uni-
fication as a general information-passing mechanism. Exam-
ple of the use of PATR-II for encoding much more complex
linguistic phenomena can be found in Shieber et al. [83].

2.2. Power: Two Variants

Augmented phrase-structure grammars such as PATR-
II can in fact be quite powerful. The ability to encode
unbounded amounts of information in the augmentations
(which PATR-II obviously allows) gives this formalism the
power of a Turing machine. As a linguistic theory, this
much power might be considered disadvantageous; as a
computer language, however, such power is clearly desir-
able, since the intent of the language is to enable the mod-
ing of many kinds of linguistic analyses from a range of
theories. As such, PATR-II is a tool, not a result.

Nevertheless, a good case could be made for maintain-
ing at least the decidability of determining whether a string
is admitted by a PATR-II grammar. This property can be
 ensured by requiring the context-free skeleton to have the
property of off-line parsability [Pereira, 83], which was used
originally in the definition of LFG to maintain the decid-
ability of that formalism [Kaplan and Bresnan, 83]. Off-line
parsability requires that the context-free “skeleton” of the
grammar allows no trivial cyclic derivations of the form
A \implies A.

2.3. Mathematical Well-Foundedness: A
Denotational Semantics

One reason for maintaining the simplicity of the bare
PATR-II formalism is to permit a clean semantics for the
language. We have provided a denotational semantics for
PATR-II [Pereira and Shieber, 84] based on the information
systems domain theory of Dana Scott [Scott, 82]. Insofar as
more complex formalisms, such as GPSG and LFG, can be
modeled as appropriate notations for PATR-II grammars,
PATR-II’s denotational semantics constitutes a framework
in which the semantics of these formalisms can also be de-
defined, discussed, and compared. As it appears that not all
the power of domain theory is needed for the semantics of
PATR-II, we are currently pursuing the possibility of build-
ing a semantics based on a less powerful model.\(^5\)

2.4. Flexibility: Modeling Linguistic Con-
structs

Clearly, the bare PATR-II formalism, as it was pre-
sented in Section 2.1, is sorely inadequate for any major
attempt at building natural-language grammars because of
its verbosity and redundancy. Efficiency of encoding was
\(^5\)But see Pereira and Shieber [84] for arguments in favor of using domain
theory even if all the available power is not utilized.
temporarily sacrificed in an attempt to keep the underlying
formalism simple, general, and semantically well-founded.
However, given a simple underlying formalism, we can build
more efficient, specialized languages on top of it, much as
MACLISP might be built on top of pure LISP. And just
as MACLISP need not be implemented (and is not imple-
mented) directly in pure LISP, specialized formalisms built
conceptually on top of pure PATR-II need not be so imple-
mented (although currently we do implement them directly
through pure PATR-II). The effectiveness of this approach
can be seen in the fact that at least a sizable portion of
English syntax has been encoded in various experimental
PATR-II grammars constructed to date. The syntactic con-
structs encoded include subcategorization of various com-
plement types (NP\(s\), SS, etc.), active, passive, “there” in-
sertion, extraposition, raising, and equi-NP constructions,
and unbounded dependencies (such as Wh-movement and
relative clauses). Other theory-dependent devices that have
been modeled with PATR-II include head-feature percola-
tion [Gazdar and Pullum, 82], and LFG-like semantic forms
[Kaplan and Bresnan, 83]. Note that none of these con-
structs and techniques required expansion of the underlying
formalism; indeed, the constructions all make use of the
techniques described in this section. See Shieber et al. [83]
for a detailed discussion of the modeling of some of these
phenomena.

The devices now available for molding PATR-II to con-
form to a particular intended usage or linguistic theory are
in their nascent stage. However, because of their great im-
portance in making the PATR-II system a usable one, we
will discuss them briefly. It is important to keep in mind
that these methods should not be considered a part of the
underlying formalism, but merely “syntactic sugar” to in-
crease the system’s utility and allow it to conform to a
user’s intentions.

2.4.1. Templates

Because so much of the information in the PATR-II
grammars under actual development tends to be encoded
in the lexicon, most of our research has been devoted to
methods for removing redundancy in the lexicon by allow-
ing the users themselves to define primitive constructs and
operations on lexical items. Primitive constructs, such as
the transitive, dyadic, or equi-NP properties of a verb, can
be defined by means of templates, that is, DAGs that en-
code some linguistically isolable portion of the DAG of a
lexical item. These template DAGs can then be combined
to build the lexical item out of the user-defined primitives.

As a simple example, we could define (with the follow-
syntax) the template Verb as

Let Verb be

<cat> = V

and the template ThirdSing as

Let ThirdSing be

<agr number> = singular
<agr person> = third

The lexical entry for “knights” would then be
The example is merely meant to be indicative of the syntax for and operation of lexical rules. We do not present this as a valid definition of Passive for any grammar we have written in PATR-II.

2.6. Implementability

Implementability is an empirical matter, given credence by the fact that we now have three implementations of the formalism. One desirable aspect of the simplicity and declarative nature of the formalism is that even though the three implementations differ substantially from one another, using different parsing algorithms (with both top down and bottom up properties), different implementations of unification, different methods of compiling the rules, all are able to run on exactly the same grammars yielding the identical results.

The three implementations of the PATR-II system currently in operation at SRI are as follows:

- An INTERLISP version for the DEC-2060 using a variant of the Cocke-Kasami-Younger parsing algorithm and the KIMMO morphological analyzer [Karttunen, 83], and a limited programming environment.
- A ZETALISP version for the Symbolics 3600 using a left-corner parsing algorithm and the KIMMO morphological analyzer, with an extensive programming environment (due primarily to Mabry Tyson) that includes incremental compilation, multiple window debugging facilities, tracing, and an integrated editor.
- A Prolog version (DEC-10 Prolog) running on the DEC-2060 by Fernando Pereira, designed primarily as a testbed for experimentation with efficient structure-sharing DAG unification algorithms, and incorporating an Earley-style parsing algorithm.

In addition, Lauri Karttunen and his students at the University of Texas have implemented a system based on PATR-II but with several interesting extensions, including disjunction and negation in the graph structures [Karttunen, 84]. These extensions will undoubtedly be integrated into the SRI systems and formal semantics for them are being pursued.

3. Conclusion

The PATR-II formalism was designed as a computer language for encoding linguistic information. The design was influenced by current theory and practice in computer science, and especially in the areas of programming language design and semantics. The formalism is simple (consisting of just one primitive operation, unification), powerful (although it can be constrained to be decidable), mathematically well-founded (with a complete denotational semantics), flexible (as demonstrated by its ability to model analyses in GPSG, LFG, DCG and other formalisms), modular (because of its higher-level notational devices such as templates and lexical rules), declarative (yielding order-independence of operations), and implementable (as demonstrated by three quite dissimilar implemented systems and one highly developed programming environment).
As we have emphasized herein, PATR-II seems to represent a convergence of techniques from several domains—computer science, programming language design, natural language processing, and linguistics. Its positioning at the center of these trends arises, however, not from the admixture of many discrete techniques, but rather from the application of a single simple yet powerful concept to the encoding of linguistic information.

References


