COMP24412, Lab 3

(Long) preamble

This one’s about grammar and parsing. I’m not asking you to write any Prolog code this time: the task this time is to add lexical entries and rules to three grammars, and to test them out using a set of Prolog programs. What you have to add will be quite small, but the grammars are quite large, and you will have to read them and understand what they say, because you won’t be able to add to them unless you understand them. The intention is that making you think about these grammars will give you some insights into the nature of natural language grammars. There is stuff to read in the specifications of the three grammars themselves: reading and assimilating that stuff is what this lab is all about. There’s a lot to read here, not all that much to do.

To load the grammars and the parser, you consult a top-level file called `setup.pl`. Once you’ve done that, doing

```prolog
?- setup(lab3a).
```

will read the grammar for exercise (a) (`lab3b`, `lab3c` for the others), and will then recompile the parser. It has to be done this way, because the parser pays attention to the sets of features used by the grammar, so you have to recompile it after you load the grammar itself. If you finish one exercise and want to move on to the next, and you do

```prolog
?- setup(lab3b).
```

without exiting from Prolog first, you’ll get messages about various things being redefined. That’s fine—we’ve changed grammars, our rules and so on are defined in a different file, so we’d expect to get warning messages.

You can then parse a piece of text by doing

```prolog
?- chart('the cat sat', P), pretty(P).
```

```
[[the, cat], sat]
P = [[the,cat],sat] ?
yes
```

(there’s not much point in the call of `pretty` here, but when they get more complicated it’s helpful) and you can get it to print out the edges that get made along the way by doing

```prolog
?- set(showSteps).
```

(you only have to do this once. If you want to stop edges being printed do `retract(showEdges)`).

The grammars all come in four sections:
Signature

A grammar file starts with a signature to say what features are being used in this grammar. This looks like

\[
\text{signature}(\text{sign}=[\text{cat}=_, \\
\text{agree}=_, \\
\text{case}=_, \\
\text{subcat}=_, \\
\text{moved}=[\text{before}=_, \text{after}=_], \\
\text{form}=_]).
\]

This one says that a \texttt{sign} has the features \texttt{cat}, \texttt{agree}, \texttt{case}, \texttt{subcat} and \texttt{moved}, and that \texttt{moved} has two subfeatures called \texttt{before} and \texttt{after}. You’ve seen all of these apart from \texttt{moved} in the lectures, and you can ignore \texttt{moved}: it’s in there because the parser is written to allow for ‘movement’, because we’re going to need that later, but for now you can just ignore it.

You can access a feature that has been specified in the current signature by writing, for instance,

\[
| \text{- cat@X -- n}.
\]

This notation is a ‘Prolog extension’: one of the advantages of languages which allow programs to be treated as data structures is that you can easily extend them, because you can transform the data structures that make up a program before they are compiled. The programs used for these labs make quite a lot of use of this idea, with the \texttt{feature@SIGN} notation being an important example.

There is another useful extension which lets us say that several items have the same value for some feature, though we may not know what that value is:

\[
| ?- \text{agree :: [X, Y, Z].}
\]

says that \texttt{X}, \texttt{Y} and \texttt{Z} have the same value for \texttt{agree}.

Types

We then get a set of ‘type definitions’ which give names to particular configurations of features. It is generally better to use these type definitions in rules and descriptions of dictionary entries than to specify values for features directly, since it makes it easier to change things later. It also makes things a bit more readable, which is nice, but the real advantage is that we can change the definition of a type once, and have that change reflected everywhere the type is used, rather than having to go and change all the rules and lexical entries that make use of it.

The type definitions are written as ordinary Prolog rules, but you can apply them using another Prolog extension. The rules below, for instance, define two simple types:
np(X) :-
    cat@X -- np.

subjCase(X) :-
    case@X -- subj.

I can use them to describe a subject-case NP by writing

subjnp(X) :-
    X <> [np, subjCase].

This rule should be read as saying that X is a subjnp if it is in the types np and subjCase. You can put as many types as you want in the list on the right-hand side of <>. If you only want one you can omit the list brackets.

Rules

Rules are as we’ve seen them in the lectures. The rule

S ==> [NP, VP] :-
    S <> s,
    NP <> [np, subjCase],
    VP <> vp,
    agree :: [NP, VP].

for instance says that you can make a sentence out of a subject-case NP and a VP if the NP and VP share the same value for agree. It is important to remember that the S, NP and VP in the rule S ==> [NP, VP] are variables, and that the type specifications S <> s and so on say what the items denoted by these variables are like. If I just wrote

S ==> [NP, VP].

I’d be saying that anything can be made out of anything followed by anything else, since I wouldn’t have said what properties these items had. NP is a variable. The Prolog interpreter does not know that I’ve used a name for this variable that indicates that I want it to denote an NP. You have to specify the required properties explicitly in the rule.
Lexicon

And then finally we get the dictionary. I’m calling the things in the dictionary ‘lexical entries’ rather than ‘words’ because for exercise (b) I’m going to want to put things in it that aren’t really words, so I’m going to use a more general term from now on. But pretty much they’re just specifications of what I know about some word (or word-part, for exercise (b)).

\[
\text{lexEntry}(i, X) : - \\
\quad X <> [\text{np, subjCase}], \\
\quad \text{agree}@X \text{ -- firstsing.}
\]

Exercises

(a) Extending the basic feature-based grammar [6 marks]

The first exercise uses a close variation on the standard feature-based grammar from the lectures. You’ll probably find something very like the solution buried in the lecture handouts. This one is really just to get you used to what the grammars look like, and what’s involved in adding things to them.

The aim of this one is to add whatever is needed to enable the system to analyse the following sentences:

(1) I know she loves me.  
(2) I gave her a present.  
(3) I gave it to her.  
(4) John loves Mary and Mary loves John.  
(5) he wore a red and black shirt.

Fix the grammar so it does not accept ‘her loves me’ or ‘I love she’.

For the first three of these, you need to extend the set of verb types (\textit{intrans}, \textit{trans}) so that they will allow for verbs that take a sentence, or two NPs, or an NP and a PP, as arguments. So you should look at the rules that mention these features, and at the lexical entries for \textit{sleep} and \textit{saw}, and try to adapt those to cover the examples.

The key to the other two is that we want ‘\textit{and}’ to link any pair of items of the same kind to make a compound of the same type – two sentences (as in (4)), two adjectives (5), two verbs (‘\textit{He buys and sells stolen cars}’, two NPs (‘\textit{I saw John and Mary}’)). You should be able to do it with
one rule, and one entry for ‘and’, but it takes a bit of thought. The standard name for words like ‘and’ (think of some others) is ‘conjunction’, so you might want to set the cat for words like this to be conj.

(b) Morphology, X-bar theory [7 marks]

The second exercise covers two topics:

i Words are made out of smaller things. Can we deal with that in the same way that sentences are made out of smaller things?

ii NPs are based on nouns, sentences are based on VPs which are based on verbs, adjectival phrases are based on adjectives, … Can we capture this notion by splitting the cat feature into two pieces—the cat itself and a level?

The signature in lab3b.pl is

\[
\text{signature}(\text{sign}=[\text{head}=[\text{cat}=\_, \text{agree}=[\text{number}=\_, \text{person}=\_], \text{case}=\_], \text{level}=\_, \text{subcat}=\_, \text{moved}=[\text{before}=\_, \text{after}=\_], \text{form}=\_]).
\]

where the cat is grouped with a number of other features that seem to be specified by the head word of the phrase but are inherited by the bigger phrase itself (so ‘sees’ is present tense and third person singular, so ‘sees the old man in the park’ is present tense and third person singular), and there is a separate feature called level.

level has several possible values—root, affix, word, 1 and 2. The first three should be reasonably self-explanatory—‘see’ is a root, ‘ing’ is an affix, ‘seeing’ is a word. The other two apply to phrases, roughly saying whether they are the largest kind of phrase that you can base on this word (level=2 or not (level=1). So ‘John loves Mary’ has cat=v, because it’s based on something verb-like (‘loves’) and bar=2 (because you can’t add anything to it), whereas ‘loves Mary’ also has cat=v, because it’s based on something verb-like (‘loves’) but has bar=1 (because you haven’t got the subject yet).

The grammar in lab3b.pl has entries for word parts, and rules for combining them, in addition to rules for combing words and phrases. So if you do

\[
\text{| ?- chart(‘the cat s walk ed’).}
\]
the system will analyse it as

[[the, [cat, s]], [[walk, ed]]]

(it won’t deal with ‘the cats walked’, because it’s not doing ‘segmentation’, so don’t try that one).

This exercise has two parts:

• add rules to cover comparative and superlative adjectives, so we can do

(6) I saw the old man.
(7) I saw the older man.
(8) I saw the oldest man.

• add rules which treat PPs as the level=2 version of prepositions, in much the same way that we’re treating sentences as the level=2 version of verbs and NPs as the level=2 version of nouns. How many analyses does your grammar assign to (10)? Some of these convey exactly the same information: pick two such analyses.

(9) I saw the man in the park.
(10) I saw the rich old man with a big nose with a spot on it in the park with a telescope.

Virtually all the work in this exercise is about understanding what’s going on. Hint: you do not need any new rules to cover the comparative and superlative forms for adjectives, just new entries in the lexicon. You will need a rule for base-form adjectives, just like we have one for base-form nouns, and you will need a rule about making PPs.

(c) Categorial grammar [7 marks]

Categorial grammar has just two rules:

\[
X_0 \Rightarrow [X_1, Y] :-
\]

\[
\begin{align*}
\text{head}_X &= \text{head}_X, \\
\text{args}_X &= [Y \mid \text{args}_X], \\
\text{dir}_Y &= \text{after}.
\end{align*}
\]

\[
X_0 \Rightarrow [Y, X_1] :-
\]

\[
\begin{align*}
\text{head}_X &= \text{head}_X, \\
\text{args}_X &= [Y \mid \text{args}_X], \\
\text{dir}_Y &= \text{before}.
\end{align*}
\]
The first says that if you’ve got an item of type $X_1$ which says that it wants a following $Y$, and there is a following $Y$, then you can combine them to form an item of the same basic kind but which no longer needs a $Y$. The second says much the same for cases where $X_1$ wanted a preceding $Y$.

When you use this kind of grammar, the definitions of individual words become much more complex. After all, most of the grammar has disappeared, or at least been hidden. The place where it’s been hidden is in the definitions of the individual words.

If you look at the lexical entries for this grammar, they look just like the ones in `lab3a.pl`, saying that ‘saw’, for instance, is a transitive verb. So where has the complexity gone? Into the definitions of things like `tverb`, where the list `args` and the direction in which they are expected to appear is specified.

Add definitions of `scompverb`, `adj` and `conj` so that this grammar can cope with (11)–(15). The adjective examples are probably the hardest bit of this: what does an adjective combine with, what is the result of the combination? You do NOT need to add new rewrite rules. Everything in categorial grammar is handled by getting the list of args right.

(11) I know she loves me.
(12) I saw the old man.
(13) I saw the rich old man.
(14) John loves Mary and Mary loves John.
(15) I wore a red and black shirt.