

1 Tokenization

2 Parse trees

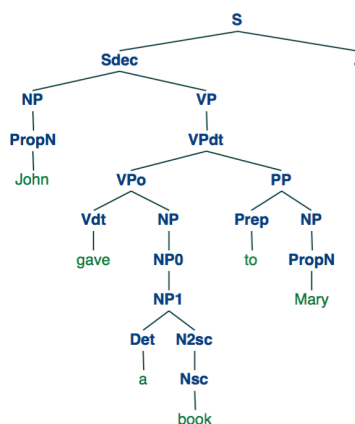


Figure 1: *John gave a book to Mary.*

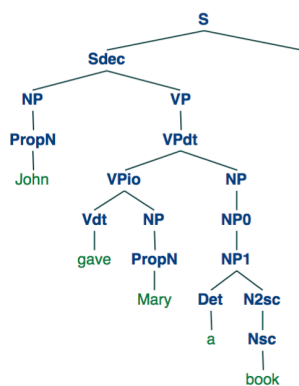


Figure 2: *John gave Mary a book.*

3 Remarks on the grammar

Design of the grammar rules

Positive aspects

Negative aspects a. Redundancy of rules in the verbal domain

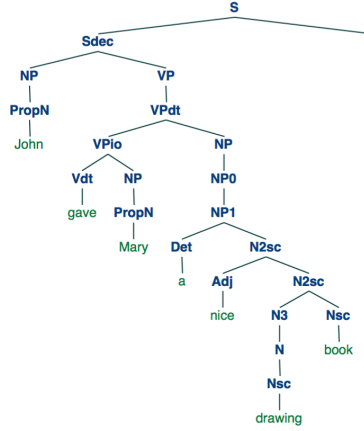


Figure 3: *John gave Mary a nice drawing book.*

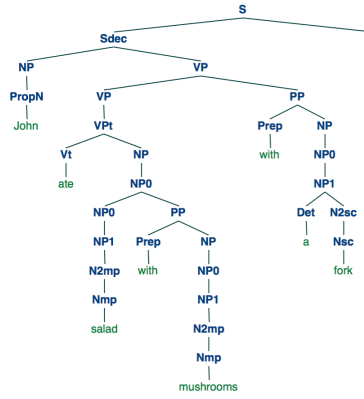


Figure 4: *John ate salad with mushrooms with a fork.*

There is a redundancy in the treatment of ditransitive verb phrases. *VPo* and *VPio* are two nodes which expand in the exact same way:

$$VPo \rightarrow Vdt \ NP \text{ and } VPio \rightarrow Vdt \ NP.$$

and therefore there is no difference between them within the grammar. From a linguistic point of view this separation might have a motivation. The ditransitive verb and its direct object are represented by *Po*, while *VPio* is a unit consisting of a ditransitive verb and its indirect object. As valid linguistic distinction as it might be, it does not need to be included in simple grammar like the one discussed here. In particular, no use is made of the distinction between that would limit overgeneration. If that was the case, then having both non-terminals would be a valid design choice. For instance, one

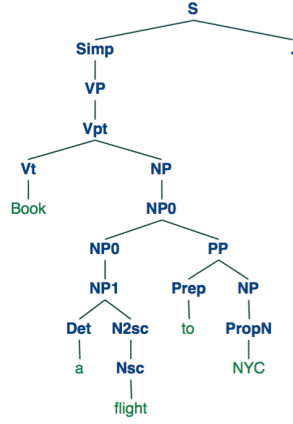


Figure 5: *Book a flight to NYC.*

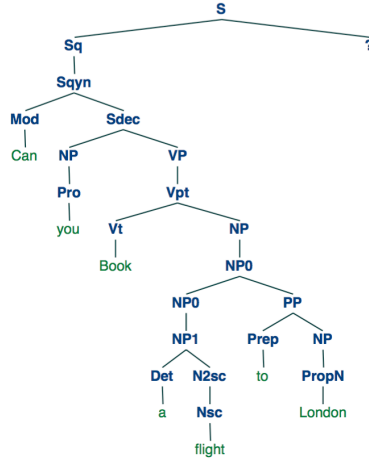


Figure 6: *Can you book a flight to London?*

might consider putting some restrictions on what kinds of NP can be direct or indirect objects. The needless distinction between VP_o and VP_{io} propagates up to the topmost VP expansion rules. As a result, there are more VP-related rules than necessary. The grammar contains

$$\begin{aligned}
 VP &\rightarrow VP_i \mid VP_t \mid VP_{dt} \mid Mod \ VP \mid VP \ Adv \mid VP \ PP \\
 VP_{dt} &\rightarrow VP_o \ PP \\
 VP_{dt} &\rightarrow VP_{io} \ NP \\
 VP_o &\rightarrow V_{dt} \ NP \\
 VP_{io} &\rightarrow V_{dt} \ NP
 \end{aligned}$$

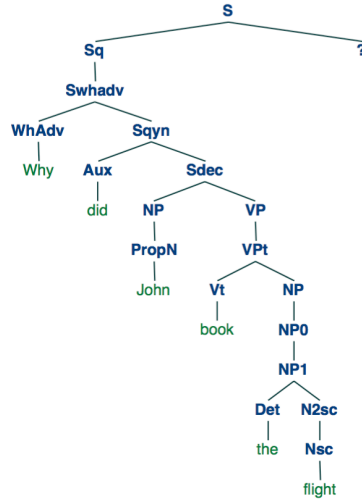


Figure 7: *Why did John book a the flight?*

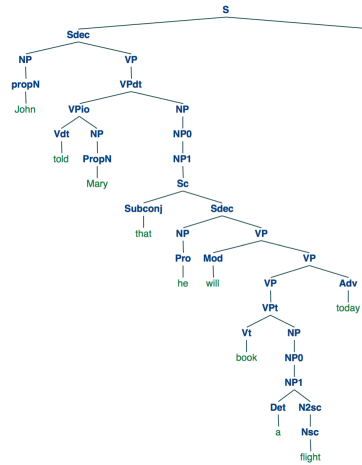


Figure 8: *John told Mary that he will book a flight today.*

but it could have contained as little as three rules instead:

$$\begin{aligned}
 VP &\rightarrow VPi \mid VPt \mid VPdt \mid Mod \ VP \mid VP \ Adv \mid VP \ PP \\
 VPdt &\rightarrow VPd \ NP \mid VPd \ PP \\
 VPd &\rightarrow Vdt \ NP
 \end{aligned}$$

b. **Redundancy of rules in the nominal domain**

Similarly, the NP0 non-terminal node seems redundant. It's purpose is to allow for recursive generation of prepositional phrases following

a noun phrase. This can be achieved by accounting for this kind of recursion within one of the existinf NP rules. We could propose:

$$\begin{aligned} NP &\rightarrow \text{PropN} \mid \text{Pro} \mid NP1 \mid NP\ PP \\ NP1 &\rightarrow \text{Det}\ N2sc \mid N2mp \mid Sc \mid NP1\ PP \end{aligned}$$

The first option has the advantage of allowing for proper names and pronouns to be modified by prepositional phrases, which could be desirable in a more comprehensive grammar, however is not necessary for the grammar at hand. The second proposition is more conservative in that it would not lead to generating any string the original grammar doesn't. The

$$NP0 \rightarrow NP1 \mid NP0\ PP$$

rule expands to an *NP1* followed by an arbitrary number of *PPs*. The same effect will be achieved by adding the *NP1* \rightarrow *NP1 PP* expansion, with the benefit of reducing the number of non-terminals. One might argue that the *NP0* node corresponds approximately to an *X'* node in X-bar syntax, and hence is theoretically justified. This, however, would need to mean that PPs to the right of the head noun are adjuncts rather than complements of that noun, which is not true. [TODO citation]

c. **Dobtfull usefulness of the *Sq* node**

The *Sq* node serves no purpose other than to reflect the conceptual relation between *Sqyn* and *Swhadv*, namely the fact both are non-terminals expanding into questions. This, however, represents only a superficial, if not mistaken, grammatical insight. In fact wh-questions and auxiliary-inversion questions are represented by very different structures (WH-phrases and Aux-phrases respectively[TODO citation]). If the motivation for including the *Sq* node is linguistic, then the reasons seem insufficient. If, on the other hand, the node was included in the grammar to simplify the expansion of *S*, then by analogy we should also include

$$Sd \rightarrow Sdec \mid Simp$$

and change the expansion of *S* to

$$S \rightarrow Sd\ '.\ ' \mid Sq\ '?\ '$$

d. **Lower case and capitalized versions of terminals.**

It does not seem necessary to include lower case and capitalized versions of the same terminals. In the sample sentences we have seen *Book*, *Can*, and *Why* being capitalized to account for their sentence-initial positions. Instead of duplicating lexical items, we could simply capitalize the first word of a generated sentence and ignore sentence-initial capitalization during parsing. There would be no ill effects of such a change, since every generated sentence is, by definition,

grammatical. Every word which happens to be initial in a sentence produced by the grammar is allowed to be there. The additional confirmation of this fact in form of the word's capitalised version being present in the grammar is unnecessary.

Overgeneration a. Lack of subject-verb agreement

There's distinction between single/count and plural/mass nouns, but no corresponding distinction in the verbal domain. The problem is somewhat masked by the fact that most of the included verbs are in past tense, where in English there are no subject-verb agreement phenomena. Nevertheless, we can see that the grammar would accept **the fork book a flight.*, a sentence which ignores verb conjugation requirements.

b. Treating VP as full sentences

The rules:

$$S \rightarrow \text{Simp} .$$

$$\text{Simp} \rightarrow \text{VP}$$

imply that any *VP* can be a grammatical sentence on its own. This, however, is true only of *VP* in imperative mood. The grammar, however, allows for treating any *VP*, e.g. **ate today*, as a full sentence. Solving this problem is not possible without substantially expanding the grammar. We would have to distinguish between infinitival and inflected forms of verbs, and extend the set of possible expansions of *VP* accordingly. We could then state that *Simp* subcategories for *VPs* whose head is an infinitival form of a verb. Therefore, we would need the following rules:

$$Vt_{inf} \rightarrow \text{Book} \mid \text{Tell} \mid$$

$$VPi_{inf} \rightarrow \text{Eat}' \mid$$

$$VPt_{inf} \rightarrow Vt_{inf} \text{ NP} \mid VPt_{inf} \text{ Adv} \mid VPt_{inf} \text{ PP}$$

$$\text{VP} \rightarrow VPt_{inf}$$

$$\text{Simp} \rightarrow VPt_{inf} \mid VPi_{inf}$$

Even then, additional restrictions would be needed to account for characteristics of the imperative mood, e.g. the fact that the second person pronoun needs to be in its reflexive form when used as a direct object of the matrix verb [TODO citation] . Otherwise the grammar would generate sentences such as **Book you a flight to London.* instead of grammatical **Book yourself a flight to London.*

c. Case agreement

The grammar does not account for case subcategorisation requirements of verbs. The only corner of the modern English syntax which still exhibits case phenomena is the pronoun system[TODO citation], however the grammar does not include a *ProNom* and a *ProAcc* preterminals. Having no distinction between nominative and accusative

forms leads to admitting non-grammatical sentences, e.g. **Mary gave **he** a nice salad.* instead of *Mary gave **him** a nice salad.*

d. **Verb and noun subcategorization for pronouns**

Verbs and nouns tend to co-occur with particular pronouns more often than with others. However, the grammar does not account for this preference. For instance, one can *eat with a fork*, *eat with Mary*, or have a *flight to London*, but the grammar also allows for having a **salad to fork* and booking a **flight with NYC*.

e. **Multiple modals**

The grammar allows for an unlimited number of modals to proceed a VP. This can lead to generating sentences such as **Can will can can John book a flight?*. However, English limits the acceptable number (up to 4, depending on dialect), orderings (evidential, epistemic, deontic), and identity (*might* or *may* being the most commonly first in the sequence (Di Paolo 1989)) of verbs in multiple modal constructions. In order to capture these restrictions we would need a whole series of serially connected expansion rules, each introducing one modal of a certain type.

4 Comments for CKY.buildIndices

```
def buildIndices(self, productions):
    """ Creates dictionaries for storing the production rules.
        In each dictionary, the rhs of a rule is the key and and
        a list of all lhs which expand as the rhs is the value. """

    # create dictionaries for unary and binary rules
    self.unary=defaultdict(list)
    self.binary=defaultdict(list)

    for production in productions:
        # separate its right hand-side from its left hand-side
        rhs=production.rhs()
        lhs=production.lhs()
        # the assumption about the rules is that rhs is non-empty
        # and rhs has no more than 2 non-terminals
        assert(len(rhs)>0 and len(rhs)<=2)
        # if the rule is unary,
        # add it's lhs to the unary dictionary under rhs key
        if len(rhs)==1:
            self.unary[rhs[0]].append(lhs)
        # if the rule is binary,
        # add it's lhs to the binary dictionary under rhs key
        # because of the assertion we know that len(rhs)==2
        else:
            self.binary[rhs].append(lhs)
```

5 Comments for CKY.unary_fill & Cell.unary_update

```
def unary_fill(self):
    """Determine the possible non-terminals that
       each terminal can result from.
       Fill the results in the middle diagonal and print"""

    for r in range(self.n-1):
        # the middle diagonal
        cell=self.matrix[r][r+1]
        # initialize the cell
        word=self.words[r]
        cell.addLabel(word)
        # recursively update the cell
        cell.unary_update(word,self.unary)
        # print out the possible non-terminals for each cell (terminal)
        if self.verbose:
            print "Unary branching rules at node (%s,%s):%s"%(r,r+1,cell.labels())

def unary_update(self,symbol,unaries):
    """Update the cell labels by adding non-terminals
       which expand as the given symbol"""

    # if the symbol is a right hand-side of a unary production rule
    if symbol in unaries:
        # add each of possible corresponding left hand-sides
        # to the cell's labels
        for parent in unaries[symbol]:
            # only add labels that is not in the cell already
            # to avoid the exponential cost of the recognition process
            if parent not in self._labels:
                self.addLabel(parent)
                # a recursive call is needed because in the grammar
                # not all unary productions are terminal productions
                self.unary_update(parent,unaries)
```

6 Comments for CKY.parse & CKY.maybe_build

```
def maybe_build(self, start, mid, end):
    """Search for the possible combinations of
       the symbols in two given cells (one from each)
       to match the rhs of binary branching rules"""

    if self.verbose:
```



```

        print "Binary branching rules for %s--%s--%s:"%(start,mid,end),

    cell=self.matrix[start][end]

    # search from the given cells
    for s1 in self.matrix[start][mid].labels():
        for s2 in self.matrix[mid][end].labels():
            # for a binary branching rule match
            if (s1,s2) in self.binary:
                # add all possible non-terminals
                # from the lhs of the rule
                for s in self.binary[(s1,s2)]:
                    cell.addLabel(s)
                # add more possible non-terminals
                # because, in the grammar, there are unary rules
                # that can produce non-terminals
                cell.unary_update(s,self.unary)
                if self.verbose:
                    print " %s -> %s %s"%(s, s1,s2),
    if self.verbose:
        print

```

*All the comments above are included in *cky.py*

7 Ambiguity Analysis

7.1 "John gave a book to Mary."

Distinct Parse(s): 3

Ambiguity Source:

7.2 "John gave Mary a book."

Distinct Parse(s): 1

Ambiguity Source:

7.3 "John gave Mary a nice drawing book."

Distinct Parse(s): 2

Ambiguity Source:

7.4 "John ate salad with mushrooms with a fork."

Distinct Parse(s): 5

Ambiguity Source:

7.5 "Book a flight to NYC."

Distinct Parse(s): 2

Ambiguity Source:

7.6 "Can you book a flight to London?"

Distinct Parse(s): 2

Ambiguity Source:

7.7 "Why did John book the flight?"

Distinct Parse(s): 1

Ambiguity Source:

7.8 "John told Mary that he will book a flight today."

Distinct Parse(s): 3

Ambiguity Source:

8 Generating a parse tree

8.1 CKY parsing

Up to now, the program is a CKY recognizer that can only determine whether a string belongs to the language generated by the given grammar. In order to extend it into a CKY parser that can construct a parse tree, back-pointers, which records the one or more ways in which a constituent spanning (i,j) can be made from constituents spanning (i,k) and (k,j) , will be attached to each label in the table cells. Therefore, after successful parsing, we can traceback from the Start Symbols (S) in the top-right-hand corner of the matrix to every terminals from the string. The final result is then a shared-forest of possible parse trees, where common tree parts are factored between the various parses [1].

8.2 Design and implementation

First, a Label class was defined to describe the labels in each cell which was represented by String. Benefited from the flexible data structure in Python, this class can be initialized with at most 4 parameters, as shown below.

List of parameters:

- **symbol** the represented label (required)
- **start** row number of the cell (optional)
- **end** column number of the cell (optional)
- **rhs** list of right-hand-sides of the production rules which generate the target symbol; mid value is also included in the list, which refers to the source cells from which the constituents of right-hand-sides came (optional)

The Label class can take different number of parameters depending on where the label came from, such as the left-hand-sides of rules and words from the input string. For instance, if a symbol was added according to a unary rule, the row number and column number will be the same as its parent label which we already knew. For the labels of words, which will be the leaf nodes of a parse tree, the value of right-hand-side will be omitted (not available).

Source Type Parameter	binary rule	unary rule	word
symbol	✓	✓	✓
start	✓		
end	✓		
rhs	✓	✓	

Table 1: Parameter Structure for Different Labels

Moreover, a number of the existing CKY and Cell methods were edited to construct/exploit this richer label structure. Three utility functions for listing, searching and updating label list are also provided.

After that, the backtrack algorithm was implemented in two functions:

- **update_trees** to update all the possible subtrees of the give label in a cell (In this implementation, the function will be ended after the first subtree was returned)
- **update_children** to construct NLTK tree nodes according to given rules

update_trees and *update_children* will call each other recursively to generate the whole parse tree.

8.3 Summary

To generate the first parse tree, we only need to pass the Start Symbols (S) in the top-right-hand corner of the matrix to *update_trees* method and it will return the result after the first parse tree was found. However, in order to return all the possible parse trees, the Start Symbols (S) in the top-right-hand corner of the matrix should be treated differently from other labels and *update_trees* method should return a list of possible subtrees to construct all the ambiguous result. The detail of generating all parse trees will not be discussed in this report.

References

- [1] B. Lang, “Recognition can be harder than parsing,” *Computational Intelligence*, vol. 10, no. 4, pp. 486–494, 1994.