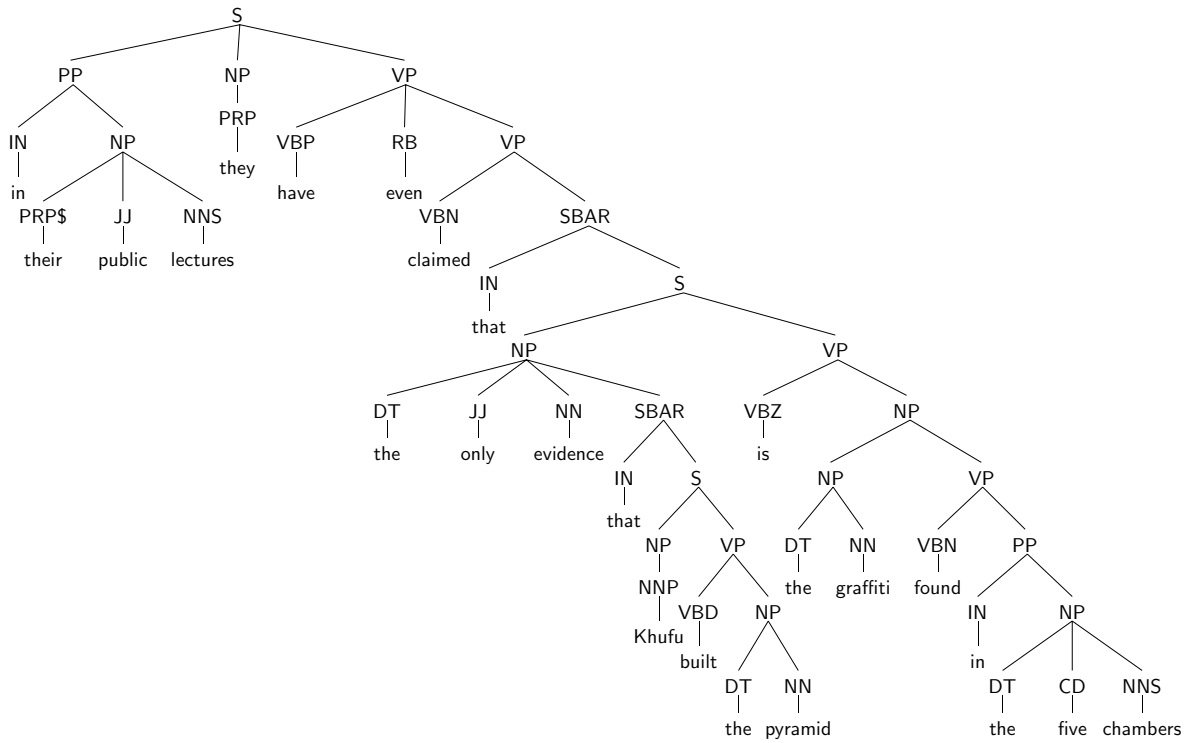


Tree Syntax of Natural Language

Lecture Note 1 for LING 4424 CS 4474 COGST 4240

Mats Rooth

1. Introduction

In linguistics and natural language processing, it is common to attribute labeled tree structures called *syntactic trees* or *parse trees* to phrases and sentences of human languages. An example is found above. The tree consists of a set of *vertices* (also known as *nodes* or *addresses*), including a unique *root vertex* which is drawn at the top. Each vertex has a label and an ordered sequence of *children*. In the example, the root vertex has label S and three children, which (in order) have labels PP, NP and VP. The child labeled VP has three children, which (in order) have the labels VBP, RB and VP. The child of VP with label VBP has one child, which has the label "have", and the vertex labeled "have" has no children. Vertices which have no childeren are called *terminal nodes*. Other nodes are *non-terminal* nodes. A vertex right above a terminal node is a *pre-terminal* node. Table 1 gives the conventional long pronunciations of the pre-terminal labels used in the example tree. These pre-terminal labels correspond to the *parts of speech* of traditional grammar. In NLP usage, the term *part of speech* is lengthened to *part of speech tag,* and shortened to *tag*. So in this tree, the tag for *built* is VBD.

TABLE 1.

label	long name	example
NN	singular noun	pyramid
NNS	plural noun	lectures
NNP	proper noun	Khufu
VBD	past tense verb	claimed
VBZ	3rd person singular present tense verb	is
VBP	non-3rd person singular present tense verb	have
VBN	past participle	found
PRP	pronoun	they
PRP$	possessive pronoun	their
JJ	adjective	public
IN	preposition	in
	complementizer	that
DT	determiner	the

Table 2 gives the other non-terminal labels in the tree. The labels ending in the letter P are known as *phrasal categories*, such as noun phrase and verb phrase. A noun phrase is, roughly speaking, a phrase organized around a noun. This noun is known as the *head* of the phrase. The head of the first NP is *lectures*, and the head of the second one is *evidence*. Similarly, a verb phrase is a phrase organized around a verb, and a prepositional phrase is is a phrase organized around a preposition.

TABLE 2.

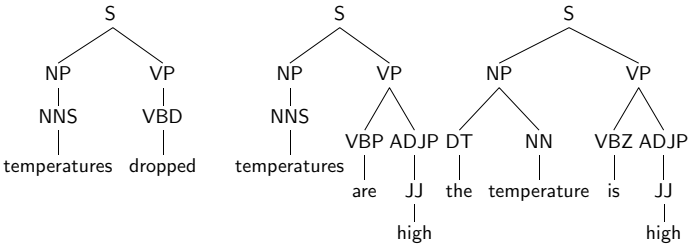
label	long name	example (represented by terminal string)
NP	noun phrase	their public lectures
VP	verb phrase	built the pyramid
PP	prepositional phrase	in the five chambers
S	sentence	Khufu built the pyramid
SBAR	sbar	that Khufu built the pyramid

2. English syntactic structures

It is useful to become familiar with the symbols used in syntactic trees, and with the tree analysis of common constructions and sentence types. In this note, we use the system of tree annotations from the Penn Treebank of English, which is a database of trees for about 50,000 English sentences. The system is on one hand a scientific hypothesis about the structure of the English language, and on the other hand an engineering standard which is used in designing and testing NLP systems. Treebanks for other languages (such as Chinese) have been published or are under development.

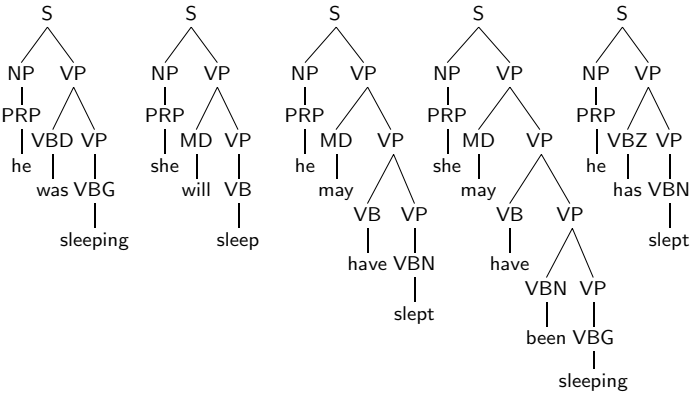
Tensed sentences and VP recursion

A minimal sentence in English consists of a subject noun phrase such as *the temperature* and a tensed verb phrase such as *dropped* or *is high.* S is the label for the sentence. In the tag for the verb heading the VP, there is a three-way distinction between past tense (tag VBD for the pre-terminal above the verb), 3rd person present tense (tag VBZ) and non-3rd person present tense (tag VBP):



This distinction is expressed in the part of speech tag, but not in the VP or S label.

Where there are auxiliary verbs (such as the modal verbs *will, can,* and *may,* or various forms of *have* and *be)*, the verbs are arrayed in a right-branching structure of VPs:



The rightmost verbs in these structures are called *main verbs*, in opposition to auxiliary verbs. However, in the Penn Treebank tag vocabulary, auxiliary verbs are not given tags different from those of main verbs, with the exception of modals and *to*.

Here is the complete vocabulary of verb tags.

TABLE 3.

Tag	Long name	Example
VBD	past tense	He ate/VBD the cookies.
		She answered/VBD the question.
VBZ	present tense	He likes/VBZ cookies.
VBP	present tense	They like/VBP cookies.
	3rd person plural	They answer/VBP such questions.
		They are/VBP tired.

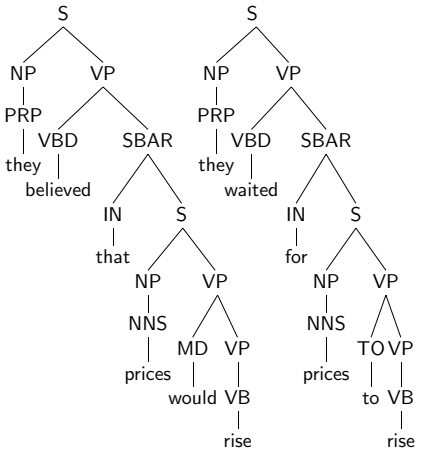
TABLE 3.

Tag	Long name	Example
VB	base	He may like/VB cookies.
		I heard her answer/VB the question.
		They may be/VB tired.
VBG	present participle, G-form	Eating/VG cookies is unhealthy.
		He likes eating/VG cookies.
VBN	past participle, N-form	He has eaten/VBN the cookies.
		She has ansered/VBN the questions.
		My question was not answered/VBN.
MD	modal	She will/MD prevail.
TO	auxiliary *to*	She expects to/TO prevail.

Most distinctions between tags correspond to overt differences in the form of the verb. VBP and VB systematically have the same form, with the exception of *are/be*.. For verbs including the most regular ones (such as *answer*), there is no distinction in form between VBD and VBN. In general, the assignment of tags is determined by context in the tree, not just by word form.

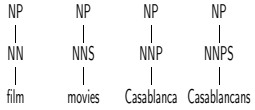
The VB form a verb is a "base" form of the verb in that, in the case of regular verbs, other forms are derived from it by adding suffixes. This process may be accompanied by minor alterations in spelling, such as consonant doubling (sit/VB, sitting/VBG) or deletion of an e (site/VB, siting/VBG). Such processes are much more elaborate in other languages

To is considered an auxiliary verb because it is found in VP recursion structures similar to what is found with modal verbs:



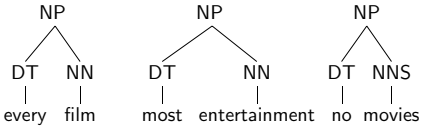
Noun phrases

A minimal noun phrase consists of just a noun:



A singular noun has tag NN, a plural noun has tag NNS, a singular proper noun has tag NNP, and a plural proper noun has tag NPS.

What is called a *determiner* may be added at the start of the noun phrase:



Some determiners can form noun phrases in isolation, with an elliptical interpretation:

Many impressed me.

Each impressed me.

Some impressed me.

*The impressed me.

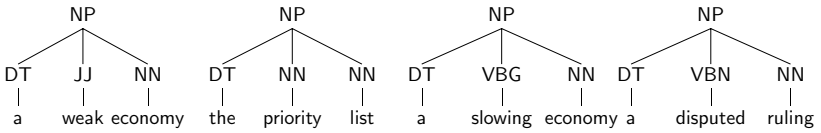
*A impressed me.

*Every impressed me.

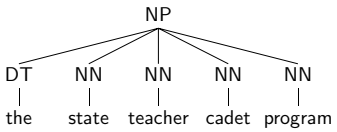
In the tree structure for these examples, an NP node dominates a DT and nothing else:

The *star* notation used above is used to mark sentences which do not sound right to the native speaker, and which, though they may possibly be comprehensible, would not be used. Such sentences are *ungrammatical* in the language under discussion. Scientific and technical work on human language takes a naturalistic view on what counts as grammatical: if a sentence sounds right to native speakers of the language, or if one can find the sentence (or a corresponding sentence pattern) being used regularly, then the sentence is considered grammatical.

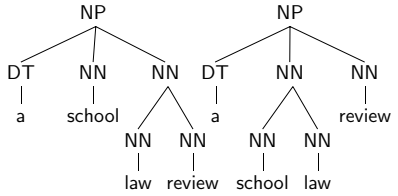
The noun in an NP can be preceded by a variety of *modifiers*, notably adjectives and other nouns, but also including G-form and N-form verbs:



Modifiers can be combined, making the NP longer:



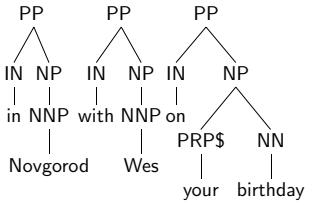
Arguably, sequences of modifiers have internal structure. There are two meanings for *school law review* (a law review at a school, and a review of school law, possibly performed in another institution such as a legislature). These correlate with two intonations (with primary strees on *law*, and primary stress on *school*, respectively). It is plausibly to attribute these different meanings and pronunciations to different tree structures, along the following lines.



As an approximation, a flat structure is used.

Prepositional phrases

A typical prepositional phrase consists of a preposition (tag IN) followed by a noun phrase. The tree structure is as follows.



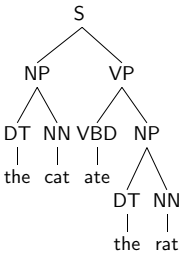
There are some systematic semantic subclasses of prepositional phrases:

TABLE 4.

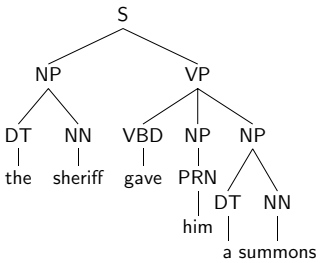
class of PPs	**examples**
temporal	on Monday, in November, after lunch
locative	in Ithaca, on campus, under the sheet
path	through downtown, into Barcelona

Complementation

A simple transitive sentence such as *the cat ate a rat* consists of a subject, a verb, and an object. The object is an NP just like the subject, and it is represented as a child of VP:

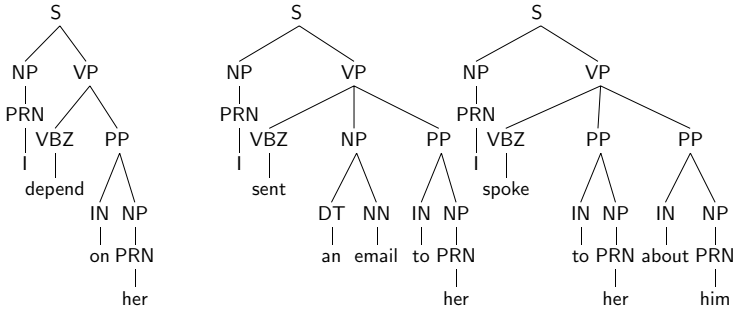


The object NP is said to be a *complement* of the verb *ate*. *Ditransitive* verbs are found with two noun phrase complements:



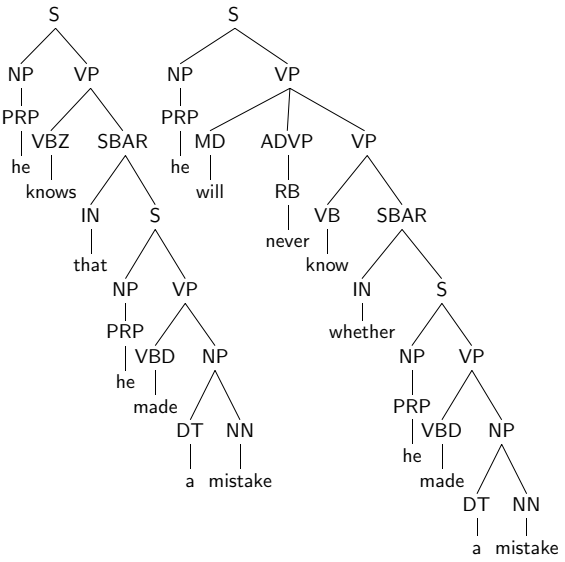
Prepositional compements

PP complements are PP children of VP, occurring alone or with another complement:

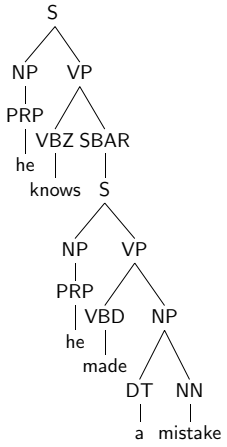


Clausal complements

Clausal complements are sentences embedded as complements of a verb. Like other complements of verbs, they are children of VP:



The label for the complement is SBAR; the SBAR begins with a *complementizer* such as *that, whether*, or *if*. The complementizers have the prepositional tag IN. The SBAR has an S child, which in these examples is a tensed sentence. Even if there is no complementizer, a SBAR node is present:



An alternative label for the complementizer is C, and an alternative label for SBAR is CP (complementizer phrase).

Selection

It is characteristic of complementation that the kind of complement which is possible correlates with the verb. If we switch the verbs in the examples above, the result is often an ungrammatical sentence:

* I depend her.

* I ate to her about him.

*He believed to her.

*He spoke whether he made a mistake.

A verb is said to *select* the complement or pattern of complements it can occur with. The complements that a verb can occur with are a property of the individual word, and this information is typically listed in a computational dictionary.

Some verbs with prepositional complements select particular prepositions:

I depend on/*in her.

He yearned for/*to an icecream cone.

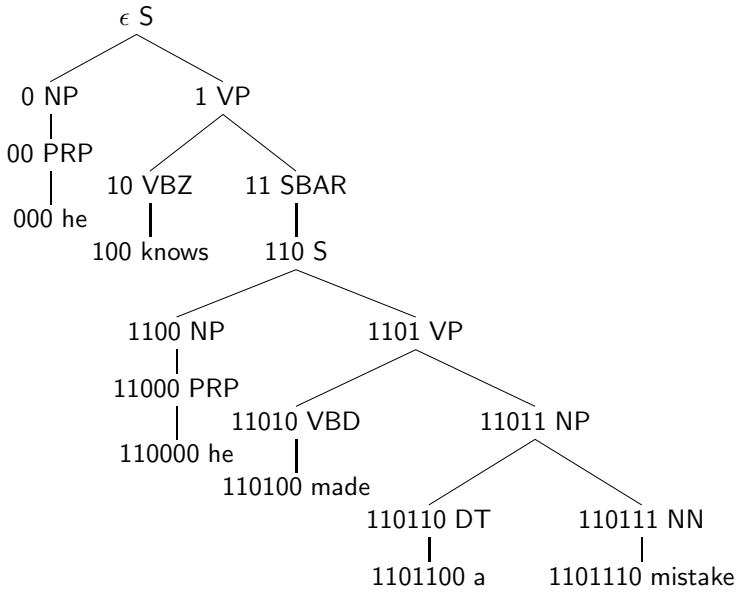
Others select a semantic class of prepositional phrases:

He left the paper in the trash. (Location)

*He left the paper into the trash. (Path)

3. Construction of trees and languages

An address of a vertex in a tree is a path specifying the route from the root to the vertex. Here is a syntactic tree drawn with addresses, in addition to the labels.



An address is a finite sequence of non-negative integers. The address of the root is the empty sequence ε. If a is an non-terminal address in the tree, then the address of the first child of *a* is *a*0. The address of the second child (if there is one) is a1, and the address of the *i*th child (counting from 0) is *ai*.

A tree domain is a tree (without labels) constructed as a set of addresses, by imposing two closure properties. In the definition, *N* is the set of natural numbers including 0, *N** is the set of finite sequences of natural numbers, including the empty sequence, the variables i and k range over N, and the variable α ranges over N*.

Definition A tree domain *X* is a subset of *N** satisfying the following conditions.

(i) If $\alpha k \in X$ and $0 \leq i < k$ then $\alpha i \in X$.

(ii) If $\alpha i \in X$ then $\alpha \in X$.

The first condition ensures that local addresses range upward from 0. The second condition ensures that an address is in the domain only if it is an extension (i.e. a child) of an address which is also in the domain.

Labels are represented as using using a function on the tree domain. Since the domain is recoverable from the function, the labeled tree can be identified with that function.

Definition A labeled tree *t* is a function such that *Dom(t)* is a tree domain.

Example The tree drawn above is the finite function *t* defined by the following table of arguments and values.

TABLE 5.

x	*t(x)*	*x*	*t(x)*	*x*	*t(x)*	*x*	*t(x)*
ε	S	10	VBZ	11000	PRP	11011	NP
0	NP	100	knows	110000	he	110110	DT
00	PRP	11	SBAR	1101	VP	1101100	a
000	he	110	S	11010	VBD	110111	NN
1	VP	1100	NP	110100	made	110111	mistake

Tree languages

In the conception adopted in formal language theory, a language is a set. For instance, the elements of the language described by the regular expression *ab** are those character strings which have length at least one and have *a* in the first position, and *b* in every other position.

Instead of strings, the elements of formal languages can be other structure objects, such as trees or graphs. For some purposes (such as a proof we will look at in a later lecture that English is not a regular language), one treats natural languages as sets of strings. In an approach which uses labeled trees to capture the syntax of human languages, it is more useful to take the elements of languages to be labeled trees. We say that the tree defined in the example above an element of a set E (English). The following tree is not an element of E.

It is clear that E is a large set. Our vocabularies consist of tens of thousands of words, which combine combinatorially, resulting in an exponential growth in the number of sentences, as sentence length increases. People often use sentences which have never been used before---I would guess that 90% of the sentences in this lecture have never been used before, even by me.

Arguably, E is an infinite set. Patterns such as the following can be extended seemingly without bound.

> This is the dog that worried the cat that chased the rat that ate the carrot that lay in the house that Jack built.

Even long sentences of this pattern are comprehensible. It seems arbitrary to say that, at some point, sentences of this pattern are no longer English sentences. If one accepts the consequence that there is no such point, then English contains a countably infinite set of sentences of this pattern.