# **Introduction and learning objectives**

So welcome to topic 6, where we're going

to talk about syntax and parsing, so let's take a look at our

course learning objectives and our topic learning objectives. So for the course learning

objectives in this particular topic, we're interested in 1, 3 and 4. So we're going to be looking at some

of the differences to rule based and statistical approaches,

with a focus definitely on the former. That's possibly what makes this topic

slightly different to some of the others. Objection number 3,

we're going to be using parsers and doing some syntactic analysis of text. So yep, certainly using some

NLP libraries for that. And objection number 4,

we'll be defining some formal grammars for fragments of natural language and looking how to apply those grammars

to do some elementary parsing. So that's it. For the course learning objectives. Let's look at the topic

learning objectives. Three of those, so we'll spend a while

understanding the fundamentals of grammar and parsing from

a conceptual point of view. Then we'll move on to implementation and we'll be applying some practical

syntax analysis techniques. And finally we'll look at a further

technique to address some of the problems with traditional grammatical formalisms. And we'll look at what's

called probabilistic parsing. So that's it. Let's get started.

# **Intro to syntax and parsing**

So in this video, we're going to start

to look in a little bit more detail at the topic of syntax and parsing. Now, you've probably heard the word syntax

and most of us have probably gone to classes at school where we've studied our

own native language or foreign languages. And we may have heard this term syntax and had a kind of intuitive

understanding of what it means. Within the context of

natural language processing, it really refers to the way

in which words are arranged. So we've actually come

across this concept before. If you think about in, right at the start

in topic one where we talked about various word sequences, we were starting to

introduce the idea that certain sequences were legitimate concerns,

other sequences, not legitimate. And in topic three when we

talked about language modeling, we saw that we could represent

word sequences in different ways, and grammar being one

typical example of that. And we also early on in the course

looked at part of speech tagging. If you recall, parts of speech are lexical

categories that we attached to words. So you can think of those as

grammatical equivalence classes, and they combine in certain orders. So there's the notion that certain

combinations are legitimate and certain are not legitimate. So at the heart of this idea of

providing structure to language at the syntactic level is

the notion of constituency. And constituency really refers to

the idea that groups of words behave as a single unit. And then a grammar is essentially

an inventory of all the constituents for a language. So, let's look at an example of that. So, to start us off let's just

look at some simple noun phrases. So up here we've got Harold the Barrel,

the Barking slugs, they and a high-class joint like Tony's,

these are all noun phrases. And one thing you notice about them

is that they can be substituted in similar syntactic environments. It actually is a little bit resonance

of the idea we had in the previous topic where we talked about word shall

be known by the company they keep. So it's a little bit similar to that

that we discussed in topic four. But look at these particular examples. We can put,

we can say Harold the Barrel cut and then finish the sentence in some way. We could say the Barking slugs are or

they sit or a high-class joint like Tony's is. And what you'll notice is that all of

these noun phrases can be followed by a verb. So that introduces the notion of

constituency that they all have this common behavior. And they can be substituted in

similar syntactic environments. So that really leads us on to the idea of

how we can model this in a formal manner. And that leads us on to what

are called context-free grammars, and these are sometimes referred to

as phrase-structure grammars. And the phrase-structure grammar

consists of a set of rules, which we'll see in a moment. And then what's called a lexicon,

which is a collection of words or symbols that exist within that language. So, if we think back to our example

from earlier, we could say, for example a noun phrase, if we look at

this little set of, these are called production rules over here on the left,

you get a symbol on the left, then you get an arrow and then you get

a collection of symbols on the right,, which we'll define in a moment. But if you think about it from the noun

phrase example that we just saw, we can say an NP, a noun phrase,

consists of a determiner, which is something like an, the or a. And then the nominal,

which behaves a little bit like a noun. Or we could say a noun

phrase could also be mapped on to what's called a proper noun. And then we can read on down the list, we could say a nominal can

be produced by a noun. And that vertical bar is an or symbol. So a noun or a nominal followed by a noun. And then the determiner can

map on to the word a or it can map on to the word the and

a noun can map on to the word flight. So these are production rules. And they're composed of

what's called non-terminals, which are things like noun-phrase,

determiner, noun, the nominal and so on. And terminal symbols, which are the actual

Words or symbols in the language. These production rules in

context-free grammars, you always have one symbol on the left,

the non-terminal, and then a combination of non-terminal

and/or terminal symbols on the right. And they express what

are called production. So you can go from the thing on

the left to the thing on the right. Now it gets interesting when we actually

start to apply these to real language because you can use them,

a little bit like the language models we saw in topic three, for

either generation to create language or to recognize the structure in language or

assign the structure. And when we're using it

in the latter manner, the structure is typically

represented as a tree, and this is called the parse tree, and

you've got an example here on the left. So we've got the sentence. And if you look at the leaves of the tree,

I shot an elephant in my pajamas. This parse tree shows how the various

constituents combine to form a higher level constituents. And they in turn combined together all

the way up to the top of the tree where we have S, which is what's referred

to as the start symbol. And in most natural language grammars, the

start symbol corresponds to a sentence but not necessarily, it can map onto

a higher level or lower level unit. And then what we refer to as

a language is the set of strings that can be generated or

are derivable from that start symbol. So let's look at some more examples. So here we've got various

production rules. You can see s, the sentence will

start symbol, maps onto noun-phrase, verb-phrase. So for example I prefer a morning flight. We've got I is the noun phrase and then

prefer morning flight is the verb phrase. And similarly on the next production

rule we've got a verb phrase, maps onto a verb and

followed by a noun phrase. So prefer a morning flight

is an example of that. Prefer is a verb and

morning flight is a noun phrase. Next one, verb phrase can be

a verb followed by a noun phrase followed by a preposition phrase. So leave is a verb noun phrase,

Boston is is a noun phrase and prepositional phrase in the morning. And then you got verb phrase

could map on to a verb followed by a prepositional phrase. So leaving on Thursday,

leaving is the verb and on Thursday is the prepositional phrase. And finally you got prepositional phrase

can be composed of a preposition followed by a noun phrase. So from is the preposition and

Los Angeles is the noun phrase. And then just to introduce

a little bit more terminology. Strings that can be derived by the grammar

are what's called grammatical. And other strings that can't be derived

by the grammar are referred to as ungrammatical. And then the process of

mapping from a string of words to a parse tree is called

syntactic parsing. And bear in mind that there are many,

many different approaches to parsing, top-down and bottom-up and

all sorts of different methods, with different strengths and

different weaknesses as well. So there's more than one way of parsing,

applying a grammar and parsing a sentence. And it's also worth knowing that,

for English at least, there are various ways you can start,

various ways you can create a sentence, there are various common

sentence structures. So there's what's called declarative,

which is essentially like saying the book is on the table,

you're declaring some propositional fact. And then there's imperative,

which is the command. So put the book on the table

would be an imperative sentence. And then there's yes-no sentences,

like is the book on the table? Would be a yes-no. And then there's wh-structures like who,

what where, when and why. So where is the book will be a

wh-structure for the sentence for English. So that's what's called

context-free grammars. And it's worth at this point just pausing

for a moment to recognize that there are other ways of thinking

about language and grammar. So whereas the previous technique,

context-free grammars or phrase structure grammars, they focus on how words

combined to form constituents and the notion of constituents

is at the heart of that. But there's other ways of

thinking about structure, syntactic structure within language. And another common formalism or approach

is what's called dependency grammar. And in contrast the phrase

structure grammar, what we're interested in here is

how words relate to each other. And that introduces the notion of what

are called dependencies where dependencies are essentially binary

asymmetric relationships. And these relationships between

what's called the head, the head of a sentence and its dependence,

and they create a labeled directed graph. Directly because the arcs

are in one direction and labeled because we label the arcs. And the head is usually a tensed verb. So here, for example, same sentence,

I shot an elephant in my pajamas. Shot is the verb, the tensed verb, so

that would be the head of this sentence. And then the arcs that you can see coming

out of that head and all the others or most of the others describe the dependency

relations going from the heads to the dependence. So shots goes to eye with

a subject relationship and also to an elephant with

an object relationship. And then to determine a modifier

from elephant to an, and a noun modifier from elephant to in. And then a prepositional

modifier from into pajamas and a determiner modifier from pajamas to my. So those are what are called

the dependency relations. And that's one other way of thinking

about grammatical structure. And it's worth noting that in English we

don't tend to have crossings that often. Because the word order is,

compared to other languages, meaningful and significant. So, like man bites dog and dog bites

man mean very, very different things. So we don't rely on the endings of words,

we rely on word order. And as a result of this,

crossings are relatively rare in English, and we call this property projective. In other words, the graph is

projective if there are no crossings. But of course, bear in mind,

in other languages, and Latin is a good example where

the word order is much more fluid. And it's all about the inflections on

the words that describe the relations that the role that they

play in the sentence. And then these crossings

are much more common. But for English at least this is

a much more typical type of analysis. So there it is,

we talked about context-free grammar and dependency grammar at a conceptual level. And in the next couple of videos we'll

have a look at how to apply these things in practice to some fragments of text.

# **Intro to syntax and parsing practice quiz**

Question 1

Which of the following is a key concept in understanding syntax?

* Dimensionality reduction
* Classification
* Constituency
* Consistency

Question 2

Context-free grammars consist of which of the following?

* Noun phrases
* A set of rules
* A lexicon
* Verb phrases

Question 3

Which of the following are legitimate sentence-level constructions in English?

* Noun phrases
* Verb phrases
* Declarative
* Imperative

Question 4

Which of the following are true of dependency grammars?

* The head is usually a tensed verb
* A graph is productive if there are no crossings
* Arcs are dependency relations
* A dependency representation is a labeled directed graph

# **Context-free grammars and recursion**

So in this segment, we're going to take

a look at some dependency parsing. Pausing and we're going to do a little

bit of practice using two software libraries to do some dependency parsing. So the first of those libraries is nltk,

and I'm sure most of you're

familiar with that by now. So we'll import nltk, and

we'll define a little grammar, and it's a dependency grammar and

a little bit like what we did earlier with the form string method

of the context free grammar. We're going to instantiate it

from the string you see here, so four lines you got shot with

dependencies on I, elephants and in and elephants dependencies on an and in, and

so an and in pajamas, pajamas to my. So that gives us seven productions,

and you can see them listed here, so the dependencies go there

from left to right, so we've got those seven dependencies. So it's not that interesting

at this point, but it starts to get interesting

when we actually apply it. That dependency grammar, we're using

a projective dependency parser, and we give it the sentence,

I shot an elephant in my pajamas. And we produce the parse trees for it, and then we iterate through those

parse trees and print them out. And you can see here, just as we saw with

the context free grammar formalism that we get this prepositional

phrase attachment ambiguity. In the first interpretation,

we've got the elephant is in the pajamas, because it's all wrapped up in this

deeply nested construction here. Where you've got the cascade bracketing

showing an elephant in our pajamas is all one, all the dependencies

all go in one direction, but here, you've got appears,

you've got elephant tan. And in, pajamas, my, at the same level,

so the in my pajamas, the term relation is with the word

the subject of the sentence I, rather than with the object

of the sentence elephant. So there you've got the same ambiguity,

but here is represented using dependency relations rather than the

constituency in the context free grammar. Now, we can do a similar thing

using another toolkit called spacy, which you may have heard of. So we'll import spacey, and

we'll instantiate our sentence before, and spacy in order to run

requires a model, english, core, web assembly which we

download from the web. And we load it and

that creates an nlp object which does all the interesting stuff, so

we apply nlp to our sentence, we get the results in

something a little called doc. And then, we go through the tokens in

the document, we print out the text, the token dependency in the next

column and an explanation for the token dependency in the third column. So you got I which is the nominal subject, shot which as we know from our earlier

video is the root of the sentence. We've got an, which is determiner,

elephant which is a direct object, in which is a preposition. My, which is a possessive or

possession modifier and pajamas which is the propositional

objects or the objects in proposition. So very useful analysis,

slightly different approach, but again, very informative and quite accurately indicates the direct

relation within this sentence. Now, we can't show

the ancestors of each term, so it's really doing the same sort

of thing that we saw earlier, going through the tokens, and

printing out the ancestors sentencies. The ancestor of I was shot, the ancestor

shot is nothing because it's the root the ancestor I is elephant and shot, and

so on down the dependency structure. Or we could do the converse, we could show the children of each term,

so I has no children, shots being the head has three,

it has I, elephant and in. An has no elephant, simply has an and

so on down the list, or you could show more of a sort of composite structure

which is the subtree for each term. So same idea, we'll just iterate

through it, and this time, instead of showing the children or

ancestors, we'll show the subtree. So I really just maps onto itself,

shots being the root, everything is in the sub tree there. I shot enough into my pajamas,

an has no subtree, but elephant has an, and elephant, and

in has in my pajamas and so on, and that's the [COUGH]

subtree representation. So there it is, that some dependency

parsing, two different approaches, one in nltk, and

one in spacy, so very useful. So have a little look at that for

yourselves and feel free to experiment.

# **Experimenting with probabilistic parsing lab**

In this exercise you will extend the probabilistic grammar shown in the video (and described in Chapter 8 of the NLTK book). Your task is to extend the coverage of the grammar to accommodate: Dative verbs, e.g. Jack gave Jill telescopes Prepositional phrases, e.g. Jack gave telescopes to Jill Hint: one solution is to introduce an intermediate rule, e.g. NP->N Adjust the rule probabilities to suit your intuitions based on how common you think certain constructions are.

# **Understanding probabilistic parsing discussion prompt**

In the previous exercise you extended the probabilistic grammar shown in the video to accommodate more complex constructions. Write a summary for each of these bullet points and post it in the discussion forum. What kinds of syntactic construction did you add to accommodate these extensions? How did you estimate their respective probabilities? How do you think such probabilities would be estimated in practice? Once you’ve posted your comments in the forum, take a look at those of other learners and comment on the differences.

# **Syntax and parsing summary**

Here we are at the

end of Topic 6, which was all about

syntax and parsing. Now let's just take

a brief moment to recap on the things

that we've covered. As you may recall, we had three topic

learning objectives. We said we'll understand

the fundamentals of grammars and parsing, and we did that from

a conceptual level. We also looked at

how we could apply syntax analysis techniques in practice with a variety

of different approaches, and we also looked at probabilistic

approaches to parsing. As you may recall, at the

start we talked about context-free grammars and the importance of

understanding constituency. We saw how we could use them for generating or

assigning structure. We talked all about production

rules and parse trees, and we could go from

a set of starting symbol S and producing an interpretation of a

sentence or an utterance, and determining whether it is grammatical or ungrammatical. That was the phrase

structure grammar approach with context-free grammar. We also looked at

dependency grammar, where instead of thinking

about how words combined, we looked at it from a different perspective of how words relate to each other. We saw how you could use the formulas like the one

shown on the screen here with binary asymmetric

relationships to create a label directed graph where the head of that graph was essentially

a tensed verb, and then you had

dependency relations going from the heads

to that dependence. Then finally, we

looked at some of the shortcomings with

context-free grammar. We saw how they create essentially equally

plausible parses and how that could lead to an

overload of possibilities. We reflected on how we as

human beings don't normally maintain dozens of different

parses in our heads, even though they are there, we tend to focus on

the most likely one. We talked about how

we would do that with attaching probabilities to the productions

and production rules, and we spent a little time showing how we can

do that in practice. You had an exercise on that

and a discussion on that. We saw the value of thinking about grammar from a

probabilistic point of view. There it is, that's

the end of Topic 6, a bit shorter than some of the others because

we know you've got your midterm

assessment coming up. Hope you found that

useful and good luck with the midterm assessment.