Automata Theory for Language

Name: Thomas Graf
Email: mathcamp@thomasgraf.net
Course Website: mathcamp.thomasgraf.net
Personal Website: thomasgraf.net

1 Language as a Mathematical Problem

Every language follows certain rules of grammar. I do not mean the usual grammar-nazi malarkey like "Do not split infinitives! Do not strand prepositions!". Rather, there are words and sentences that are correct, and others that have something wrong with them.

Example 1

Consider the English word *denaturalization*. It is built up from discrete parts:

1. nature
2. nature + al = natural
3. natural + ize = naturalize
4. de + naturalize = denaturalize
5. denaturalize + ation = denaturalization

No other order of these parts produces a good word of English:
denaturizalation, *ationalizenaturde*, *naturalizedeation*, ...

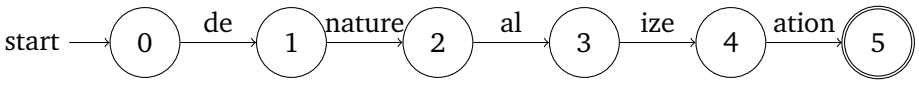
[Exercise 1] How many combinations are mathematically possible?

We see, then, that there must be a very strict system of rules that determines the order in which parts of a word may appear. What can we say about this rule system? Can we study it just like, say, the laws that determine how atoms combine into molecules?

One strategy would be to look at the human brain. After all, language must involve some kind of brain mechanisms that allow us to produce correct forms while avoiding incorrect ones. But research in *neurolinguistics* has shown that this approach won't get us very far — the brain is just too complicated to be studied this way. Instead, we need a more abstract model of these brain mechanisms. And math allows us to do just that!

2 Graphs for Language: Finite-State Automata

Our mathematical model are *finite-state automata* (FSA), which are a special case of labeled graphs. Here is a simple example of an automaton representing *denaturalization*:



- The circles are the vertices of the graph, which are called *states*.
- The *arcs* connecting the states all must have a label.
- The *start* arrow indicates the beginning point of the automaton, called the *initial state*.
- The doubly circled state is a *final state*. It marks the end point of the automaton.

An FSA can be used to succinctly describe words and sentences. Every path through the automaton that takes you from an initial state to a final state represents a well-formed structure. In the automaton above, there is only one such path, which takes us from 0 to 1, then to 2, 3, 4, and finally 5. But automata can have multiple paths through them.

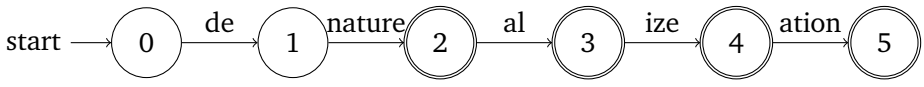
Example 2

Besides *denaturalization*, English also has the words

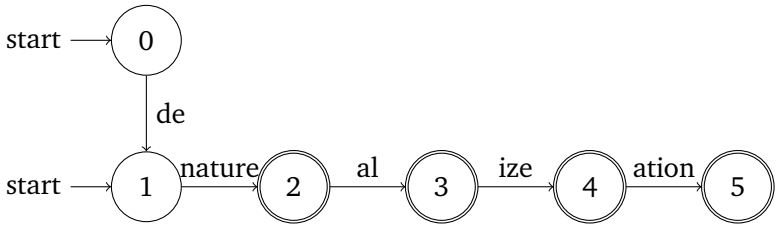
- *nature*,
- *natural*,
- *naturalize*,
- *denaturalize*,
- *naturalization*.

We can modify the previous automaton so that it also allows for these other forms. Let's do this step by step.

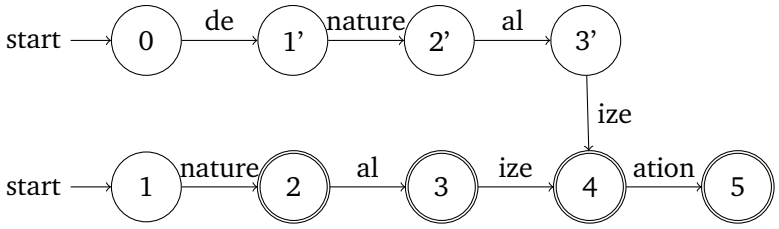
First we want to allow words to also end with *nature*, *al*, and *ize*, so we make the states that are reached through those arcs final.



But every path must still start with *de*, so the automaton does not allow the patterns *nature*, *natural*, or *naturalize*. To fix this, we make 1 an initial state, too.



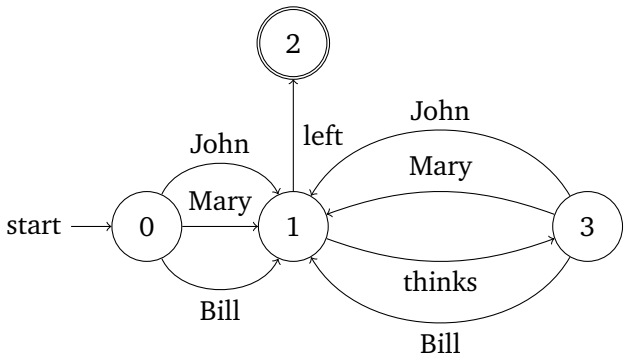
But this automaton is too general, it also allows patterns like *denature* or *denatural*. This is a little tricker to fix: we add a few more states to keep track of the fact that after *de*, the first final state can only be one reached by *ize*.



When you now try all possible paths from an initial state to a final state, you'll see that they are exactly

- *nature*,
- *natural*,
- *naturalize*,
- *denaturalize*,
- *naturalization*,
- *denaturalization*.

FSAs can be much more complex than this. As long as the number of states is finite, pretty much any graph you can imagine is an FSA. You can have multiple arcs leaving a node, multiple arcs leading to the same node, or arcs that lead back to a previous node, possibly even the one they came from (*loops*). That is to say, you can have convoluted automata like the one below:



[Exercise 2] What collection of sentences is described by the automaton above?

3 Some Curious Limits of Language

Linguists put the parts that make up *denaturalization* into three distinct groups:

stem the base form, which can appear by itself; *nature*

prefix a part that can only appear before a stem; *de-*

suffix a part that can only appear after a stem; *-al*, *-ize*, *-ation*

[Exercise 3] Prefixes and suffixes can sometimes be iterated. Your grandfather's father is your *great grandfather*, whose father is your *great great grandfather*, whose father is your *great great great grandfather*, and so on. Try to write an FSA that produces *grandfather*, *great grandfather*, and so on. You may analyze *grandfather* as a stem.

[Exercise 4] Of course we also have *great grandmother*, *great great grandmother*, and so on. How could the automaton from the previous example be modified to also allow these words?

Those are not the only classes — among other things there are also *circumfixes*. An example of a circumfix is German *ge- -t*, which is used to form the participle form of a verb. So *rauf-en* 'brawl-to' can be turned into *ge-rauf-t* '(has) brawled'.

[Exercise 5] German actually has a split with respect to this circumfix. For some verbs it is *ge- -t*, for others it is *ge- -en*. And for some verbs, the form is irregular.

Verb stem	Infinitival form	Past participle form
lauf 'run'	laufen	gelaufen
rauf 'brawl'	raufen	gerauft
sauf 'guzzle'	saufen	gesoffen
tauf 'baptize'	taufen	getauft

Write an FSA that produces all the correct infinitival and past participle forms, and only those. Try to keep the number of states as small as possible.

One would expect that languages can freely choose whether they use prefixes, suffixes or circumfixes. But this does not seem to be the case. What we find is that across languages, circumfixes do not seem to be iterable.

Example 3 Saying 'the day after tomorrow' in German and Ilocano

German has a much simpler way than English to say 'the day after tomorrow'. You just take *morgen* (the word for 'tomorrow') and add the prefix *über* 'over'. What makes this even nicer is that *über* can be iterated.

1. morgen 'tomorrow'
2. über-morgen 'the day after tomorrow'
3. über-über-morgen 'the day after the day after tomorrow'
4. ...

Ilocano (an Austronesian language spoken on the Philippines) also has a single word for 'the day after tomorrow', but it is built with the circumfix *ka- -an* instead of a prefix.

1. bigat 'tomorrow'
2. ka-bigat-an 'the day after tomorrow'

In contrast to German *über-*, however, Ilocano *ka- -an* cannot be freely iterated. So you cannot say something like *ka-ka-bigat-an-an*.

[Exercise 6] It is very easy to give an FSA for the German pattern — it's just a variation of the *great* ⋯ *great grandfather* pattern we saw before for English. Similarly, an automaton for the Ilocano pattern is easy to come up with. But what if Ilocano were also unbounded, so that we could also have words like *ka-ka-ka-bigat-an-an-an*, i.e. words with the same number of *ka-* prefixes and *-an* suffixes. Can you write an automaton for this?

The difference between German *über-* and Ilocano *ka- -an* is an interesting puzzle, and it is part of a much larger observation: prefixes and suffixes can be freely iterated, but circumfixes cannot. Why should languages work this way? FSAs provide an answer.

4 The Limits of Finite-State Automata (aka the Hard Part)

FSAs are fairly powerful devices, and they have found many applications in the real world. They control elevators, help biologists with genome sequencing, give you word suggestions when you write a text message, and are even involved in the automatic creation of subtitles for Youtube videos. Tons of technology uses FSAs, including language technology.

But FSAs cannot do everything. There is a very strict bound on the complexity of the patterns they can handle. And what we will see now is that even though FSAs can iterate prefixes and suffixes (as you've already shown yourself with the FSA for *great* ⋯ *great grandfather*), they cannot correctly iterate circumfixes.

Before we state the theorem, let us think about how FSAs actually work. In particular, what do the states stand for? Suppose you have taken a path that leads to state 1 in the convoluted automaton from the previous section. How can you continue to get a well-formed pattern? Well, by taking any path that leads you from 1 to a final state. There may be multiple paths to choose from, but anyone will do. It does not

matter how you actually got into 1. Whether you came from 0 to 1 via *John*, *Mary*, or *Bill* is completely irrelevant, the only thing that matters is that you are in state 1 now and may take any path that gets you from 1 to 2.

Crucial Insight 1 If you have paths that take you from an initial state to the same state, then those paths can be continued in exactly the same fashion towards a final state.

FSAs have only finitely many states. Suppose we have some pattern, and we compare all paths that could possibly arise in that pattern. We put two paths in the same bin if they can be continued in exactly the same fashion. Then the pattern can be handled by an FSA only if we end up with a finite number of distinct bins.

Example 4

Bins for the *great* ⋯ *great grandfather* Pattern Consider the pattern *grandfather*, *great grandfather*, *great great grandfather*, and so on. If the pattern starts with *grandfather*, then there is nothing else that can be added, so there are no continuation paths at all. For *great*, one possible continuation path is *grandfather*, producing *great grandfather*. But we could also continue with *great grandfather*, *great great grandfather*, and so on. Here's a table:

Incoming Path	**Continuation Paths**
grandfather	—
great	grandfather, great grandfather, …
great great	grandfather, great grandfather, …
great great great	grandfather, great grandfather, …
⋮	

So there's only two bins for paths. One is for *grandfather*, which cannot be continued. The other one is for any path that starts with *great*, as they can all be continued by an arbitrary number of *great*s followed by *grandfather*. Since we only have two bins, the pattern can be handled by an FSA.

The bins are essentially the states in the automaton: two strings are in the same bin because they take you to the same state in the automaton, and from there you can proceed in a specific manner irrespective of how you got there. But a finite-state automaton can only have a finite number of states, so it can only describe patterns where the number of distinct bins is finite.

Crucial Insight 2 A pattern can be captured by an FSA only if its paths can be grouped into finitely many bins.

This insight allows us to show that unbounded circumfixation is impossible with FSAs. The proof is a joint class exercise.

5 Wrapping Up

The iterability difference between prefixes and suffixes on the one hand and circumfixes on the other is really puzzling from non-mathematical perspective. But once we look

at them from a mathematical perspective, we see that there is a huge complexity difference: FSAs can iterate prefixes and suffixes, but not circumfixes. So if the mechanisms our brain uses to handle word structure are similar to FSAs, then it is not surprising that circumfixes are never iterated in any languages — the relevant mechanisms simply cannot do it!

We have accomplished something that few people would consider possible: we have studied language from a mathematical perspective, and we were able to explain properties that are shared by all human languages by purely mathematical means.

The Take Home Message. Mathematics explains language!

[Exercise 7] Word structure indeed seems to be fully captured by FSAs. But there are ways to show that the structure of English sentences is more complicated then what FSAs can handle. Do you have an idea what the argument might be?

Hint: The sentences below play a crucial role.

- A man left.
- A man that a woman saw left.
- A man that a woman that a woman saw saw left.
- A man that a woman that a woman that a woman saw saw saw left.

6 I Wanna Learn More!

Cool, here's a few pointers. I am also working on an interactive learning platform that teaches mathematical techniques for linguistics, and **I am looking for volunteers** to test it and provide feedback. If you are interested, or if you have any questions on today's material, shoot me a line: mail@thomasgraf.net

- **Automata Theory**
  - *Introduction to the Theory of Computation*
    undergraduate course at Stony Brook, Computer Science
  - Michael Sipser. 2012. 3rd edition.
    Introduction to the Theory of Computation. (yep, that's the same name)

    Pro tip: The 2nd edition is almost the same and used copies are much cheaper. A decent scan can easily be found on the first Google page, I'm not sure about its legality.
- **Computational Linguistics**
  - *North American Computational Linguistics Olympiad*
    we have training sessions at Stony Brook
  - *Language and Technology*
    undergraduate course at Stony Brook, Linguistics
  - *Computational Linguistics*
    undergraduate course at Stony Brook, Linguistics

  - *Natural Language Processing*
    undergraduate course at Stony Brook, Computer Science
  - Markus Dickinson, Chris Brew, and Detmar Meurers. 2012.
    Language and Computers.

    Used copies are available for 10 bucks on Amazon.

- **Language from a Formal and Cognitive Perspective**
  - Check out my website, teaching materials can be found under *Students*
  - Steven Pinker. *The Language Instinct: How the Mind Creates Language*.

- **The Human Mind as a Computer**
  - Douglas Hofstadter. *Gödel, Escher, Bach: An Eternal Golden Braid*.

Group Exercise

The Ilocano pattern is:

ka	bigat	an
kaka	bigat	anan
kakaka	bigat	ananan
kakakaka	bigat	anananan
	⋮	

We can write this more succinctly as ka^n bigat an^n $(n \geq 0)$.
Suppose the word has already started with the following string:

What are the possible continuation paths?

Example. If the word starts with *ka*, the continuation paths are

- bigat an,
- ka bigat an an,
- ka ka bigat an an an,
- ...