NLP overview

Natural Language Processing

Piotr Fulmański

FACULTY OF MATHEMATICS
AND COMPUTER SCIENCE
University of Lodz

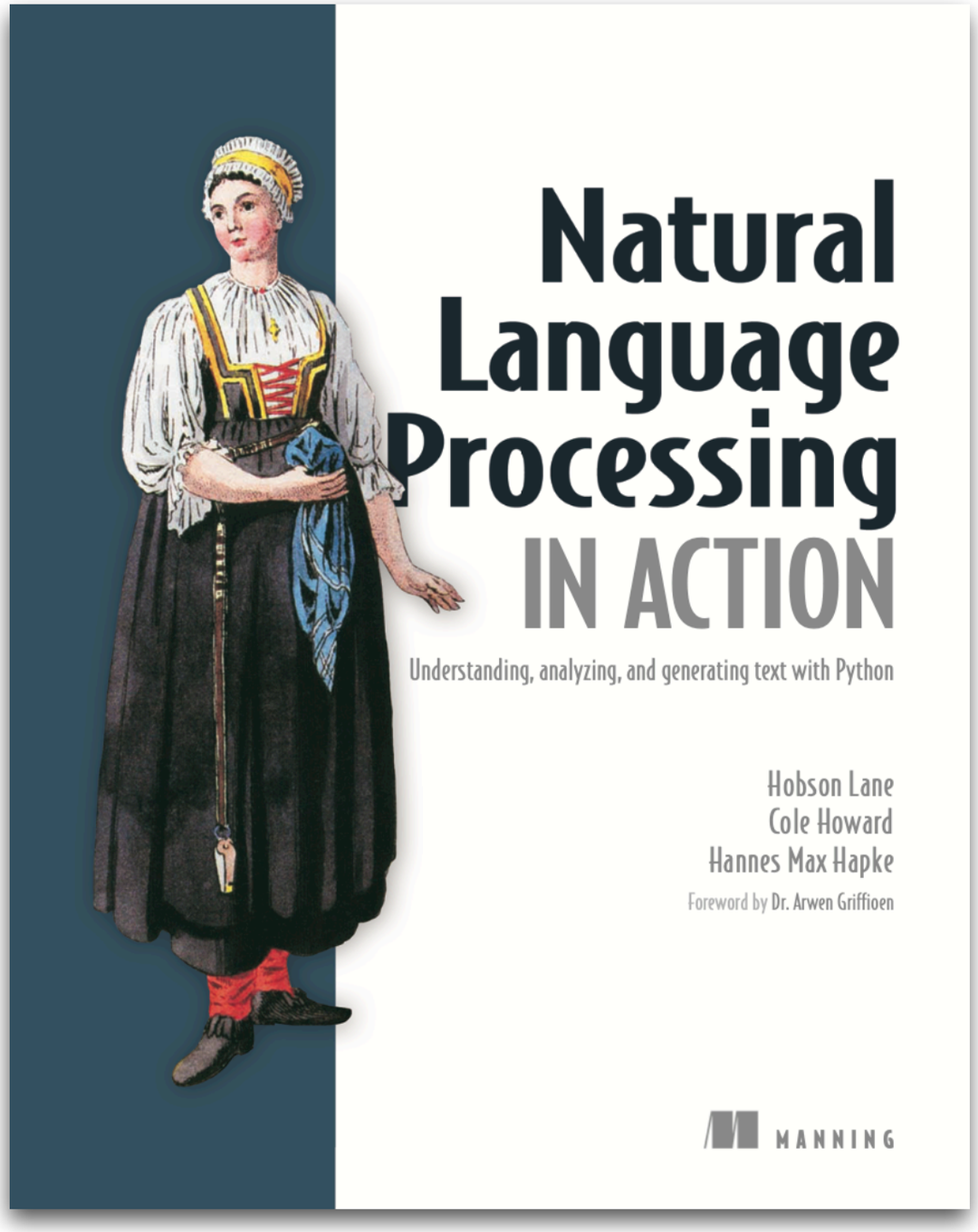
Lecture goals

- Languages and their grammar
- General introduction to language processing
- Language processing with FSM (Finite State Machine)
- Language of characters sequences - simple "chatbot"

Natural Language Processing in Action

by Hobson Lane
Cole Howard
Hannes Max Hapke

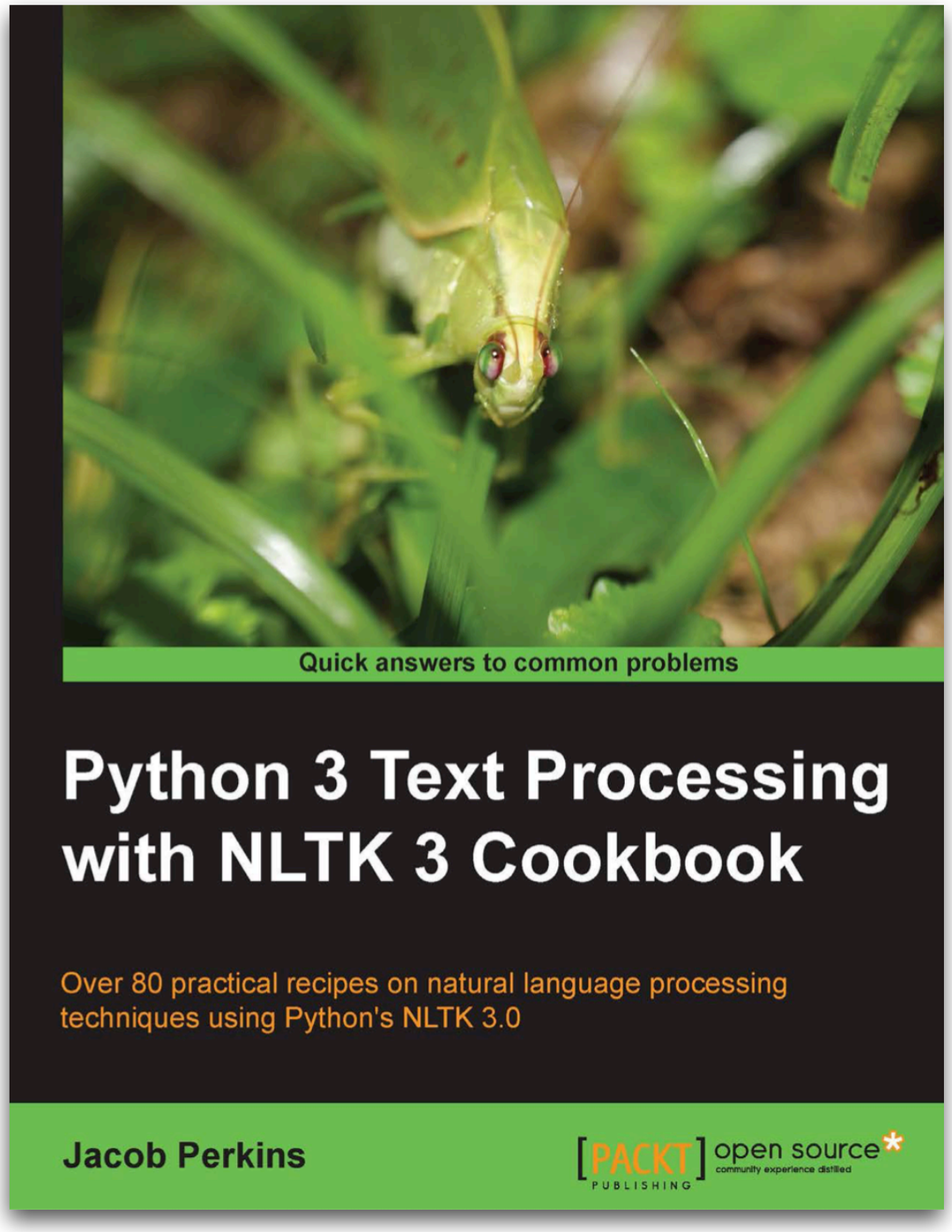
Manning Publications, 2019



Python 3 Text Processing with NLTK 3 Cookbook

by Jacob Perkins

Packt Publishing, 2014



Languages and their grammar

FORMAL GRAMMAR AND LANGUAGES

Languages and their grammar

FORMAL GRAMMAR AND LANGUAGES

See another presentation.

Natural language vs. programming language

Natural language vs. programming language

- Machines have been processing languages since computers were invented. However, these “formal” languages—such as early languages Ada, COBOL, and Fortran—were designed to be interpreted (or compiled) only one correct way.

Natural language vs. programming language

- Machines have been processing languages since computers were invented. However, these “formal” languages—such as early languages Ada, COBOL, and Fortran—were designed to be interpreted (or compiled) only one correct way.

- A computer program written with a programming language tells a machine exactly what to do.

Natural language vs. programming language

Natural language vs. programming language

- Natural languages aren’t intended to be translated into a finite set of mathematical operations, like programming languages are.

Natural language vs. programming language

- Natural languages aren't intended to be translated into a finite set of mathematical operations, like programming languages are.
- The word "natural" in "natural language" is used in the same sense that it is used in "natural world." Natural, evolved things in the world about us are different from mechanical, artificial things designed and built by humans.

Natural language vs. programming language

Natural language vs. programming language

Natural languages **can’t be directly translated** into a precise set of mathematical operations, **but they do contain information and instructions** that can be **extracted**. Those pieces of information and instruction can be stored, indexed, searched, or immediately acted upon.

Natural language

English? Polish?

We focus entirely on text documents and messages, not spoken statements. English? Polish?

Natural language

English? Polish?

We focus entirely on text documents and messages, not spoken statements. English? Polish?

EN:
two, second

Natural language

English? Polish?

We focus entirely on text documents and messages, not spoken statements. English? Polish?

EN:
two, second

PL:
dwa, dwie, dwoje, dwóch, dwaj, dwiema, dwóm, dwoma, dwojga, dwojgu, dwojgiem, dwójka, dwójki, dwójkę, dwójką, dwójce, dwójko

Natural language

English? Polish?

We focus entirely on text documents and messages, not spoken statements. English? Polish?

Natural language

English? Polish?

We focus entirely on text documents and messages, not spoken statements. English? Polish?

These functionality questions start by exploring what the user will ask of the system, or what problems they have they will want the system to solve for them.

Natural language

English? Polish?

We focus entirely on text documents and messages, not spoken statements. English? Polish?

Natural language

English? Polish?

We focus entirely on text documents and messages, not spoken statements. English? Polish?

Use of inclusive language

Natural language

English? Polish?

We focus entirely on text documents and messages, not spoken statements. English? Polish?

Use of inclusive language

Inclusive language acknowledges diversity, conveys respect to all people, is sensitive to differences, and promotes equal opportunities. Articles should make no assumptions about the beliefs or commitments of any reader, should contain nothing which might imply that one individual is superior to another on the grounds of race, sex, culture or any other characteristic, and should use inclusive language throughout. Authors should ensure that writing is free from bias, for instance by using 'he or she', 'his/her' instead of 'he' or 'his', and by making use of job titles that are free of stereotyping (e.g. 'chairperson' instead of 'chairman' and 'flight attendant' instead of 'stewardess').

Natural language

English? Polish?

We focus entirely on text documents and messages, not spoken statements. English? Polish?

Use of inclusive language

Inclusive language acknowledges diversity, conveys respect to all people, is sensitive to differences, and promotes equal opportunities. Articles should make no assumptions about the beliefs or commitments of any reader, should contain nothing which might imply that one individual is superior to another on the grounds of race, sex, culture or any other characteristic, and should use inclusive language throughout. Authors should ensure that writing is free from bias, for instance by using 'he or she', 'his/her' instead of 'he' or 'his', and by making use of job titles that are free of stereotyping (e.g. 'chairperson' instead of 'chairman' and 'flight attendant' instead of 'stewardess').

From:
https://www.elsevier.com/journals/neural-networks/0893-6080/guide-for-authors

Natural language

English? Polish?

We focus entirely on text documents and messages, not spoken statements. English? Polish?

Natural language

English? Polish?

We focus entirely on text documents and messages, not spoken statements. English? Polish?

When we say “good morning”, we assume that you have some knowledge about what makes up a morning, including not only that mornings come before noons and afternoons and evenings but also after midnights. And you need to know they can represent times of day as well as general experiences of a period of time. The interpreter is assumed to know that “good morning” is a common greeting that doesn’t contain much information at all about the morning. Rather it reflects the state of mind of the speaker and her readiness to speak with others.

Natural language

English? Polish?

Natural language

English? Polish?

- Dzień dobry - powiedział Bilbo [...].

Natural language

English? Polish?

- Dzień dobry - powiedział Bilbo [...].

- Co chcesz przez to powiedzieć? - spytał [Gandalf]. - Czy życzysz mi dobrego dnia; czy oznajmiasz, że dzień jest dobry, niezależnie od tego, co ja o nim myślę; czy sam się dobrze tego ranka czujesz, czy może uważasz, że dzisiaj należy być dobrym?

Natural language

English? Polish?

Natural language

English? Polish?

Współczesna nowomowa:

Natural language

English? Polish?

Współczesna nowomowa:

- Corpo-language: "zanim będziesz sendowała", "hasło mi expirowało", "to jest taki chalange dla niej", "zrobił krótki research.

Natural language

English? Polish?

Współczesna nowomowa:

- Corpo-language: "zanim będziesz sendowała", "hasło mi expirowało", "to jest taki chalange dla niej", "zrobił krótki research.

- Język młodzieżowy.

Natural language

English? Polish?

Współczesna nowomowa:

- Corpo-language: "zanim będziesz sendowała", "hasło mi expirowało", "to jest taki chalange dla niej", "zrobił krótki research.
- Język młodzieżowy.
- Krystyna Chodorowska Kre(jz)olka (w: Nowa Fantastyka 3’2014)

Natural language

English? Polish?

We focus entirely on text documents and messages, not spoken statements.
English? Polish?

Natural language

English? Polish?

We focus entirely on text documents and messages, not spoken statements. English? Polish?

So theory of mind about the human language processing is based on one a powerful assumption: we have an access to a lifetime of common sense knowledge about the world. Thanks this we are able to say a lot with just few words. Another point of view is that this implicit assumption often leads to many misunderstandings.

Natural language

English? Polish?

We focus entirely on text documents and messages, not spoken statements. English? Polish?

So theory of mind about the human language processing is based on one a powerful assumption: we have an access to a lifetime of common sense knowledge about the world. Thanks this we are able to say a lot with just few words. Another point of view is that this implicit assumption often leads to many misunderstandings.

There is no clear “theory of mind” you can point to in an NLP pipeline. However, we can build ontologies, or knowledge bases, of common sense knowledge to help interpret machines statements that rely on this knowledge.

Language processing

So extracting information isn’t at all like building a programming language compiler. The most promising techniques bypass the rigid rules of regular grammars (patterns) or formal languages.

Language processing

Whenever we type a text a computer sees only a sequence of numbers which at the lowest level is a sequence of 0’s and 1’s: “01000111 01101111 01101111 …”.

Language processing

Whenever we type a text a computer sees only a sequence of numbers which at the lowest level is a sequence of 0's and 1's: "01000111 01101111 01101111 ...".

- How can we program a chatbot to respond to this binary stream intelligently?

Language processing

Whenever we type a text a computer sees only a sequence of numbers which at the lowest level is a sequence of 0's and 1's: "01000111 01101111 01101111 ...".

- How can we program a chatbot to respond to this binary stream intelligently?
- Could a nested tree of conditionals (`if-else` statements) check each one of those bits and act on them individually?

Language processing

This would be equivalent to writing a special kind of program called a *finite state machine* (FSM). And this is one possible approach to NLP: **the pattern-based approach**.

Language processing

VERY SIMPLE PATTERN-BASED LANGUAGE

Language of characters sequences.

It is used in safe or lock.

With this language we can "tell" a "sentence" which, when correctly „understood", is used to unlock protected things."

Language processing

VERY SIMPLE PATTERN-BASED LANGUAGE

Language of characters sequences.

Language processing

VERY SIMPLE PATTERN-BASED LANGUAGE

Language of characters sequences.

1. One predefined sequence, like "*12345*", "*abcd*", „*1A3b*”.

Language processing

VERY SIMPLE PATTERN-BASED LANGUAGE

Language of characters sequences.

1. One predefined sequence, like "*12345*", "*abcd*", „*1A3b*”.
2. Sequence as a pattern:
   three digits, then two to four letters but no "greater" than "g" and finally digit.

Language processing

VERY SIMPLE PATTERN-BASED LANGUAGE

Language of characters sequences.

1. One predefined sequence, like "*12345*", "*abcd*", „*1A3b*”.
2. Sequence as a pattern:
   three digits, then two to four letters but no "greater" than "g" and finally digit.
3. Sequence as a complicated pattern:
   three digits, then two to four letters but no "greater" than "g" and finally digit which is equal to the number of letters used before.

Language processing

VERY SIMPLE PATTERN-BASED LANGUAGE

Language of characters sequences.

Language processing

VERY SIMPLE PATTERN-BASED LANGUAGE

Language of characters sequences.

Case 1: One predefined sequence, like "*12345*", "*abcd*", „*1A3b*”.

Language processing

VERY SIMPLE PATTERN-BASED LANGUAGE

Language of characters sequences.

Case 1: One predefined sequence, like "*12345*", "*abcd*", „*1A3b*”.

Implementation: hardcoded string.

Language processing

VERY SIMPLE PATTERN-BASED LANGUAGE

Language of characters sequences.

Language processing

VERY SIMPLE PATTERN-BASED LANGUAGE

Language of characters sequences.

Case 2: Sequence as a pattern:
three digits, then two to four letters but no "greater" than "g" and finally digit.

Language processing

VERY SIMPLE PATTERN-BASED LANGUAGE

Language of characters sequences.

Case 2: Sequence as a pattern:
three digits, then two to four letters but no "greater" than "g" and finally digit.

Implementation: regular expressions.

Language processing

VERY SIMPLE PATTERN-BASED LANGUAGE

Language of characters sequences.

Language processing

VERY SIMPLE PATTERN-BASED LANGUAGE

Language of characters sequences.

Case 3: Sequence as a complicated pattern:
three digits, then two to four letters but no "greater" than "g" and finally digit which is equal to the number of letters used before.

Language processing

VERY SIMPLE PATTERN-BASED LANGUAGE

Language of characters sequences.

Case 3: Sequence as a complicated pattern:
three digits, then two to four letters but no "greater" than "g" and finally digit which is equal to the number of letters used before.

Implementation: FSM

Language processing

FSM

FSM is the most general approach.

Language processing

FSM

FSM is the most general approach.

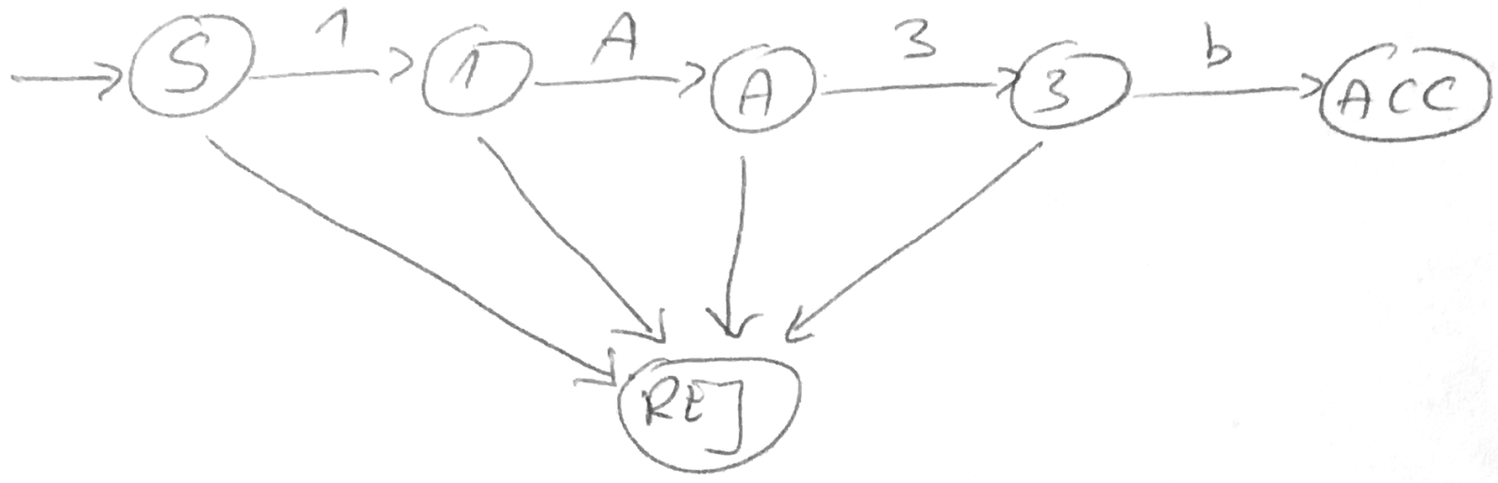
FSM for case 1:

Language processing

FSM

FSM is the most general approach.

FSM for case 1:



Language processing

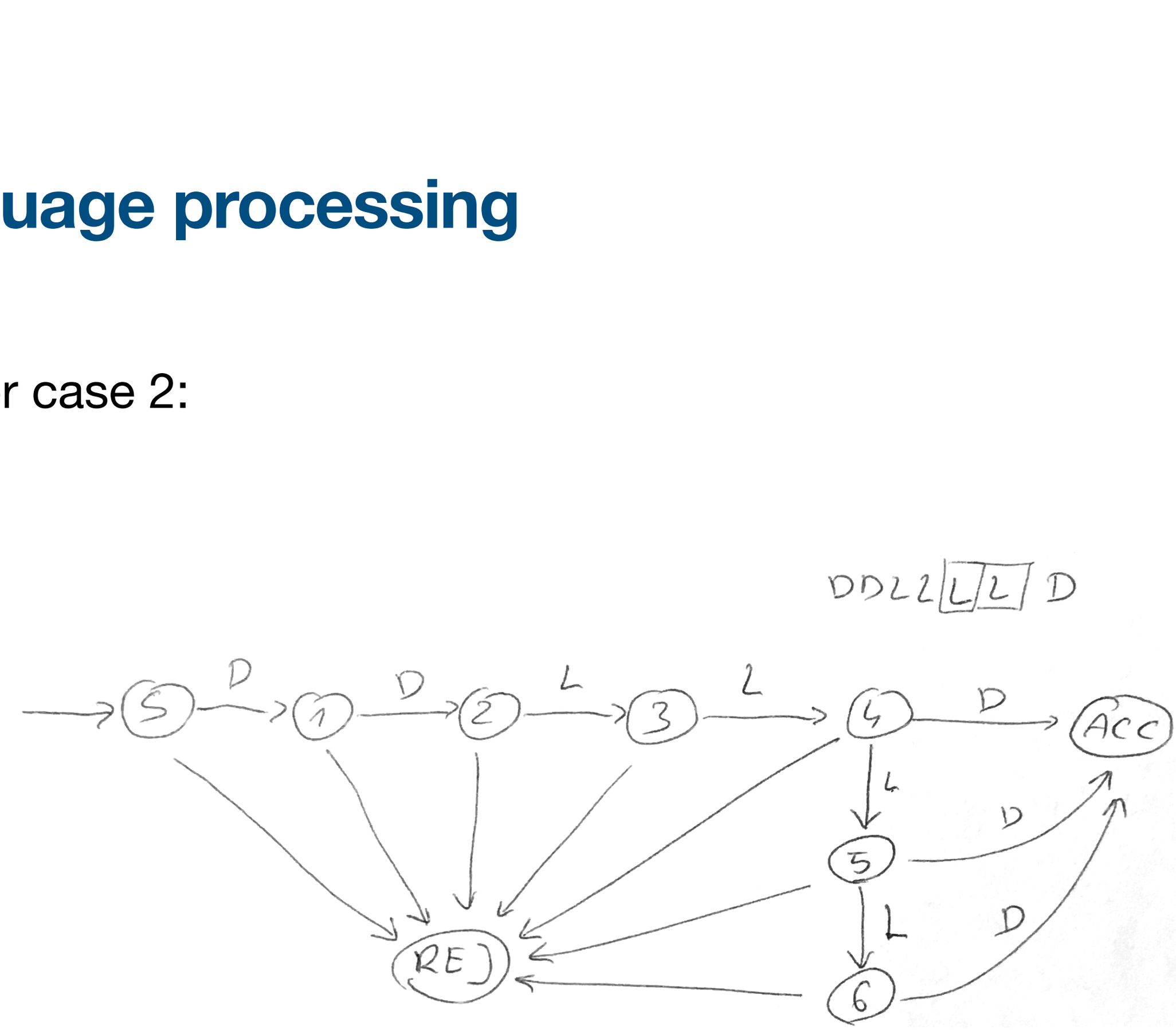
FSM

FSM for case 2:

Language processing

FSM

FSM for case 2:



Language processing

FSM

FSM for case 3:

Language processing

FSM PSEUDOCODE

Language processing

FSM PSEUDOCODE

```
routes = [{"currentState": ..., "event": ..., "newState": ...},      <-- route number 1
          ...
          {"currentState": ..., "event": ..., "newState": ...}]      <-- route number N
```

Language processing

FSM PSEUDOCODE

Language processing

FSM PSEUDOCODE

```
inputString = "..."
currentState = START
i = 0

while (currentState != ACCEPT) {
  c = inputString[i]
  event = getEventCode(currentState, c)
```

```
if (action == REJECT) {
  currentState = REJECT
  break
}
```

Language processing

FSM PSEUDOCODE

```
inputString = "..."
currentState = START
i = 0

while (currentState != ACCEPT) {
  c = inputString[i]
  event = getEventCode(currentState, c)

if (action == REJECT) {
  currentState = REJECT
  break
}


noRoute = TRUE

for route in routes {
  if (currentState == route["currentState"] AND action == route["event"]) {
    currentState == route["newState"]
    noRoute = FALSE
    break
  }
}

if (currentState == REJECT || noRoute == TRUE) {
  break
}
```

Language processing

FSM PSEUDOCODE

```
inputString = "..."
currentState = START
i = 0

while (currentState != ACCEPT) {
  c = inputString[i]
  event = getEventCode(currentState, c)

  if (action == REJECT) {
    currentState = REJECT
    break
  }


  noRoute = TRUE

  for route in routes {
    if (currentState == route["currentState"] AND action == route["event"]) {
      currentState == route["newState"]
      noRoute = FALSE
      break
    }
  }

  if (currentState == REJECT || noRoute == TRUE) {
    break
  }


  i = i + 1
}
```

Language processing

CHATBOT WITH LANGUAGE OF CHARACTERS SEQUENCES

Language processing

CHATBOT WITH LANGUAGE OF CHARACTERS SEQUENCES

Although this number sequences language is one of the simplest one, it’s not so simple that we can’t use it in a chatbot. We can use it to recognize a key phrase or command and "unlock" a particular action or behavior.

Languages and their grammar

CHATBOT WITH LANGUAGE OF CHARACTERS SEQUENCES

Languages and their grammar

CHATBOT WITH LANGUAGE OF CHARACTERS SEQUENCES

- Any math equation or programming language expression is an example of a formal language statement.

Languages and their grammar

CHATBOT WITH LANGUAGE OF CHARACTERS SEQUENCES

- Any math equation or programming language expression is an example of a formal language statement.
- Formal languages are a subset of natural languages.

Languages and their grammar

CHATBOT WITH LANGUAGE OF CHARACTERS SEQUENCES

- Any math equation or programming language expression is an example of a formal language statement.
- Formal languages are a subset of natural languages.
- Many natural language statements can be matched or generated using a formal language grammar, like regular expressions.

Languages and their grammar

CHATBOT WITH LANGUAGE OF CHARACTERS SEQUENCES

- Any math equation or programming language expression is an example of a formal language statement.
- Formal languages are a subset of natural languages.
- Many natural language statements can be matched or generated using a formal language grammar, like regular expressions.
- That’s the reason we talked about FSM.

Languages and their grammar

CHATBOT WITH LANGUAGE OF CHARACTERS SEQUENCES

Languages and their grammar

CHATBOT WITH LANGUAGE OF CHARACTERS SEQUENCES

Chatboot based game:

Languages and their grammar

CHATBOT WITH LANGUAGE OF CHARACTERS SEQUENCES

Chatboot based game:

- A Small Talk At The Back Of Beyond,
  https://scriptwelder.itch.io/a-small-talk

Bibliography

- [Lan] Hobson Lane, Cole Howard, Hannes Max Hapke, *Natural Language Processing in Action*, Manning Publications, 2019
- [Per] Jacob Perkins, *Python 3 Text Processing with NLTK 3 Cookbook*, Packt Publishing, 2014