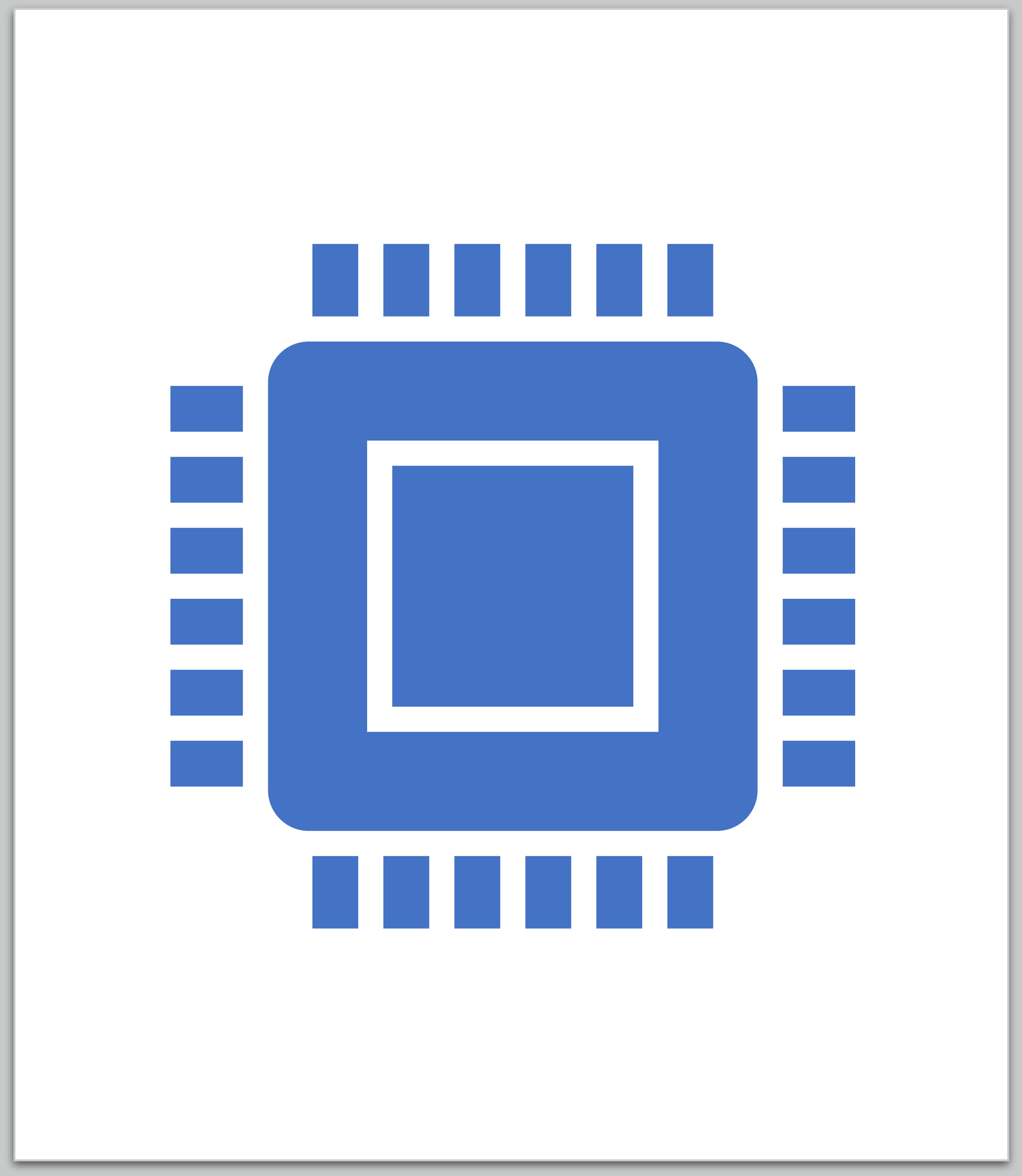
NLP - Basic units of processing

Liad Magen – Asigmo Spring 2022



Basic units of processing

What do we process?

- A Collection of Documents

  By date or domain
- Document
- Section

  Subsection
- Paragraph
- Sentence

  Phrases
- Word

  Characters / Morphemes

What do we process?

- A Collection of Documents

  By date or domain

- **Document**
- Section

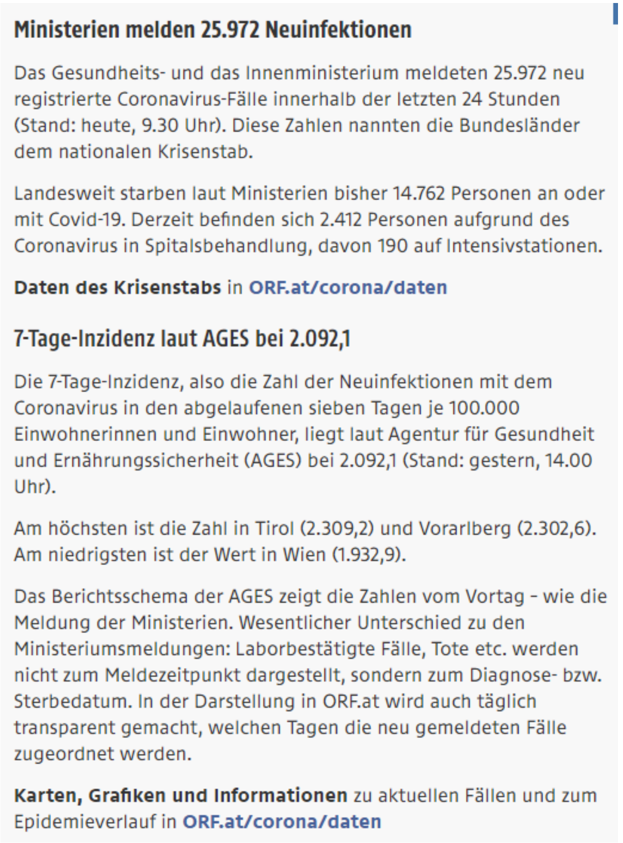
  Subsection

- Paragraph
- Sentence

  Phrases

- Word

  Characters / Morphemes



Ministerien melden 25.972 Neuinfektionen

Das Gesundheits- und das Innenministerium meldeten 25.972 neu registrierte Coronavirus-Fälle innerhalb der letzten 24 Stunden (Stand: heute, 9.30 Uhr). Diese Zahlen nannten die Bundesländer dem nationalen Krisenstab.

Landesweit starben laut Ministerien bisher 14.762 Personen an oder mit Covid-19. Derzeit befinden sich 2.412 Personen aufgrund des Coronavirus in Spitalsbehandlung, davon 190 auf Intensivstationen.

Daten des Krisenstabs in **ORF.at/corona/daten**

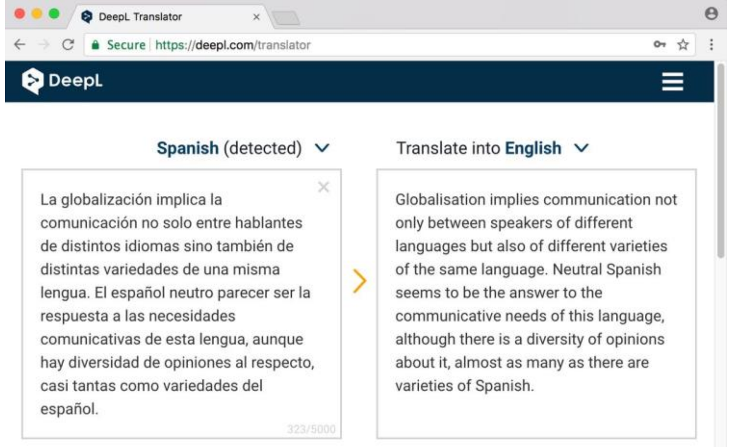
7-Tage-Inzidenz laut AGES bei 2.092,1

Die 7-Tage-Inzidenz, also die Zahl der Neuinfektionen mit dem Coronavirus in den abgelaufenen sieben Tagen je 100.000 Einwohnerinnen und Einwohner, liegt laut Agentur für Gesundheit und Ernährungssicherheit (AGES) bei 2.092,1 (Stand: gestern, 14.00 Uhr).

Am höchsten ist die Zahl in Tirol (2.309,2) und Vorarlberg (2.302,6). Am niedrigsten ist der Wert in Wien (1.932,9).

Das Berichtsschema der AGES zeigt die Zahlen vom Vortag - wie die Meldung der Ministerien. Wesentlicher Unterschied zu den Ministeriumsmeldungen: Laborbestätigte Fälle, Tote etc. werden nicht zum Meldezeitpunkt dargestellt, sondern zum Diagnose- bzw. Sterbedatum. In der Darstellung in ORF.at wird auch täglich transparent gemacht, welchen Tagen die neu gemeldeten Fälle zugeordnet werden.

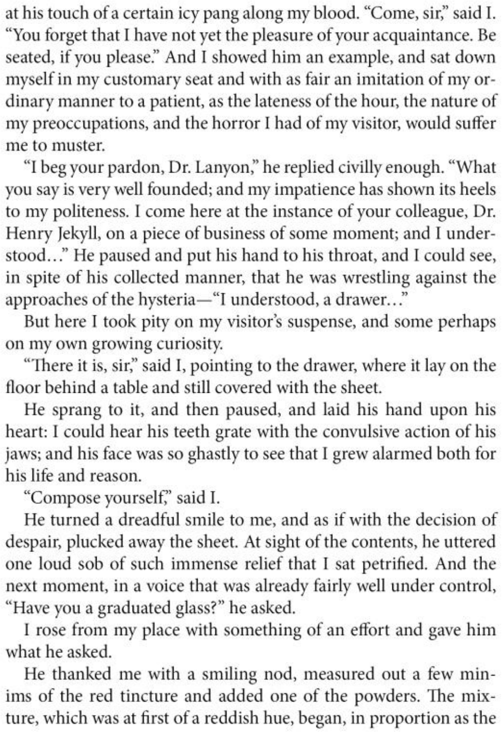
Karten, Grafiken und Informationen zu aktuellen Fällen und zum Epidemieverlauf in **ORF.at/corona/daten**



La globalización implica la comunicación no solo entre hablantes de distintos idiomas sino también de distintas variedades de una misma lengua. El español neutro parecer ser la respuesta a las necesidades comunicativas de esta lengua, aunque hay diversidad de opiniones al respecto, casi tantas como variedades del español.

323/5000

Globalisation implies communication not only between speakers of different languages but also of different varieties of the same language. Neutral Spanish seems to be the answer to the communicative needs of this language, although there is a diversity of opinions about it, almost as many as there are varieties of Spanish.



at his touch of a certain icy pang along my blood. "Come, sir," said I. "You forget that I have not yet the pleasure of your acquaintance. Be seated, if you please." And I showed him an example, and sat down myself in my customary seat and with as fair an imitation of my ordinary manner to a patient, as the lateness of the hour, the nature of my preoccupations, and the horror I had of my visitor, would suffer me to muster.

"I beg your pardon, Dr. Lanyon," he replied civilly enough. "What you say is very well founded; and my impatience has shown its heels to my politeness. I come here at the instance of your colleague, Dr. Henry Jekyll, on a piece of business of some moment; and I understood…" He paused and put his hand to his throat, and I could see, in spite of his collected manner, that he was wrestling against the approaches of the hysteria—"I understood, a drawer…"

But here I took pity on my visitor's suspense, and some perhaps on my own growing curiosity.

"There it is, sir," said I, pointing to the drawer, where it lay on the floor behind a table and still covered with the sheet.

He sprang to it, and then paused, and laid his hand upon his heart: I could hear his teeth grate with the convulsive action of his jaws; and his face was so ghastly to see that I grew alarmed both for his life and reason.

"Compose yourself," said I.

He turned a dreadful smile to me, and as if with the decision of despair, plucked away the sheet. At sight of the contents, he uttered one loud sob of such immense relief that I sat petrified. And the next moment, in a voice that was already fairly well under control, "Have you a graduated glass?" he asked.

I rose from my place with something of an effort and gave him what he asked.

He thanked me with a smiling nod, measured out a few minims of the red tincture and added one of the powders. The mixture, which was at first of a reddish hue, began, in proportion as the

Document Processing

What do we process?

- A Collection of Documents

  By date or domain
- Document
- Section

  Subsection
- Paragraph
- **Sentence**

  Phrases
- Word

  Characters / Morphemes

What is a sentence?

Python:

```
sentences = text.split(".")
```

Full stop, period.

Sentence can also end with: **! : ; ? …**

Numbers: 0.99 // 1.234,56

Date : 28.02.2022

A '.' Can also represent part of a sentence:

- www.url.com
- I.B.M // E.E.G

What is a word?

Sequence of characters?

Basic unit of meaning?

White-space **tokenization**?

What is a word? (Clitics)

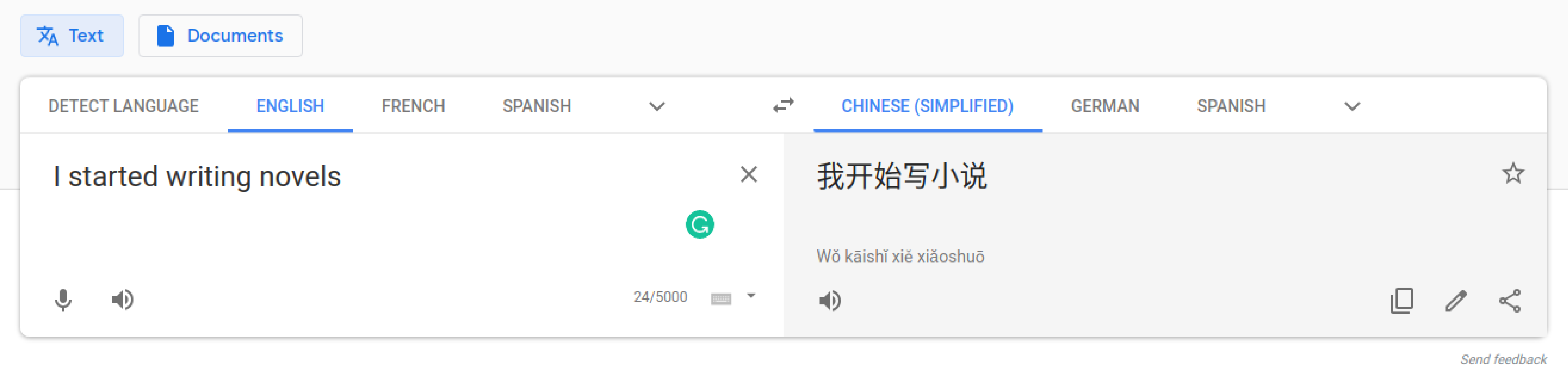
There's

Doesn't

John's

unlucky

Chinese has no spaces!



I started writing novels

我开始写小说

Wǒ kāishǐ xiě xiǎoshuō

Compounds & Orthography considered as one word

- Ice cream
- Web site
- New York
- New York-based

Phrases

- Gave up
- Gave away
- Took a picture
- Took a brake
- Took a risk
- Took the toy apart

- German: Schalten Sie es ein (➔ einschalten)

Reflexive verbs // Verb conjunction

- **French**:
  Tu te llaves
- **Portuguese**:
  Sente-se
- **Spanish**:
  Recuérdame / le dijeron
- **Albanian**:
  Laj duart

  Noun conjunction: duar -> duart

Case study:

Sentence Boundary Detection

Question:

When do we need (or don't need) to divide a document into sentences?

Not needed:

if it's already divided; Search Query; Working on a document-level (e.g., document classification)

What about translation – needed or not?

Anaphora / Cataphora / Co-reference resolution

Human: What do we want?
Computer: Natural-Language Processing!

Human: When do we want it?
Computer: When do we want what?

Sentence Boundary Detection

Python:

```
sentences = text.split(".")
```

Sentence Boundary Detection

However, if you picked up a certain book from the library, you can opt to start up your own biplane instead of boarding the zeppelin and forgo the hassle of having to navigate around the blimp’s infrastructure. Flying the plane requires you to utilize the number pad (in a slightly less frustrating way than the fights) to dodge the Luftwaffe assault while Henry blasts the enemy planes. Depending on how many planes you destroy before you get shot down, you will have to face fewer roadblocks on the way to Alexandretta when you continue the rest of the journey by car. The guards you stop you at each roadblock act as in Brunwald; i.e., find out a way to talk or bribe your way around them, or just get to punching their faces in. A certain item can even help you skip over every single guard just by showing it to them!

Sentence Boundary Detection

However, if you picked up a certain book from the library, you can opt to start up your own biplane instead of boarding the zeppelin and forgo the hassle of having to navigate around the blimp’s infrastructure. Flying the plane requires you to utilize the number pad (in a slightly less frustrating way than the fights) to dodge the Luftwaffe assault while Henry blasts the enemy planes. Depending on how many planes you destroy before you get shot down, you will have to face fewer roadblocks on the way to Alexandretta when you continue the rest of the journey by car. The guards you stop you at each roadblock act as in Brunwald**; i.e.,** find out a way to talk or bribe your way around them, or just get to punching their faces in. A certain item can even help you skip over every single guard just by showing it to them!

Sentence Boundary Detection

Python:

```
sentences = text.split(". ")
```

Sentence Boundary Detection

WASHINGTON — Former President George W. Bush called on Americans on Saturday to put aside partisan differences, heed the guidance of medical professionals and show empathy for those stricken by the coronavirus and the resulting economic devastation.

In a three-minute video message, Mr. Bush, who rarely speaks out on current events, struck a tone of unity that contrasted with the more combative approach taken at times by President Trump as the former president evoked the sense of national solidarity in the wake of the attacks of Sept. 11, 2001.

Sentence Boundary Detection

Learning system, that **learns**, according to text features, which ones are attributed to sentence ends.

- Q: How to get annotated data?
- Q: Which kind of data to use?
- Q: Which clues could hint of end-of-sentences?

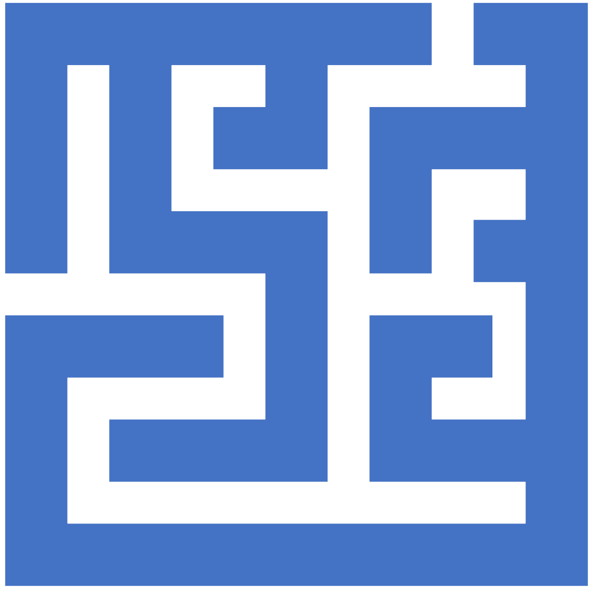
Clues to identify end-of-sentence

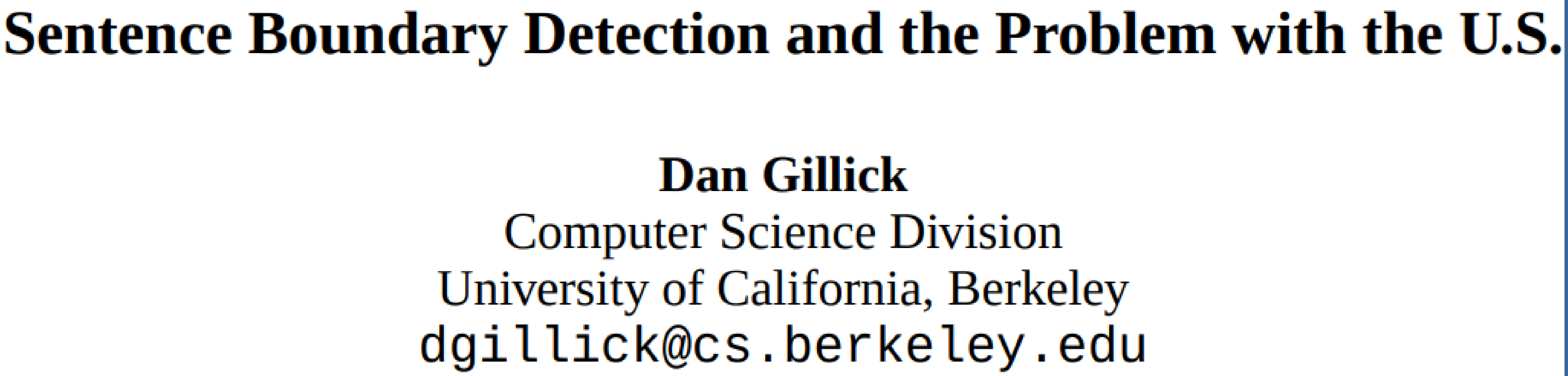
- A dictionary of safe-words:
  Mr. Dr. Ltd. Inc. U.S. e.g. i.e.
- Hinting words in a windowed neighborhood
- An ensemble or a combination of clues:
  - *Capital letter after ". "*
    Though not perfect: *George W. Bush*
- Word length: short words are probably part of a word

ML Task: Weighting the hints as a learnable solution.

Weighting the hints as a learnable solution

- Collect the data
- Annotate
- Define the 'hints' (a.k.a. *feature engineering*)
- Identify the problem: Binary classification
  - Decide which of the hints to use to determine end-of-sentence
- Train
- Evaluate





Sentence Boundary Detection and the Problem with the U.S.

Dan Gillick
Computer Science Division
University of California, Berkeley
dgillick@cs.berkeley.edu

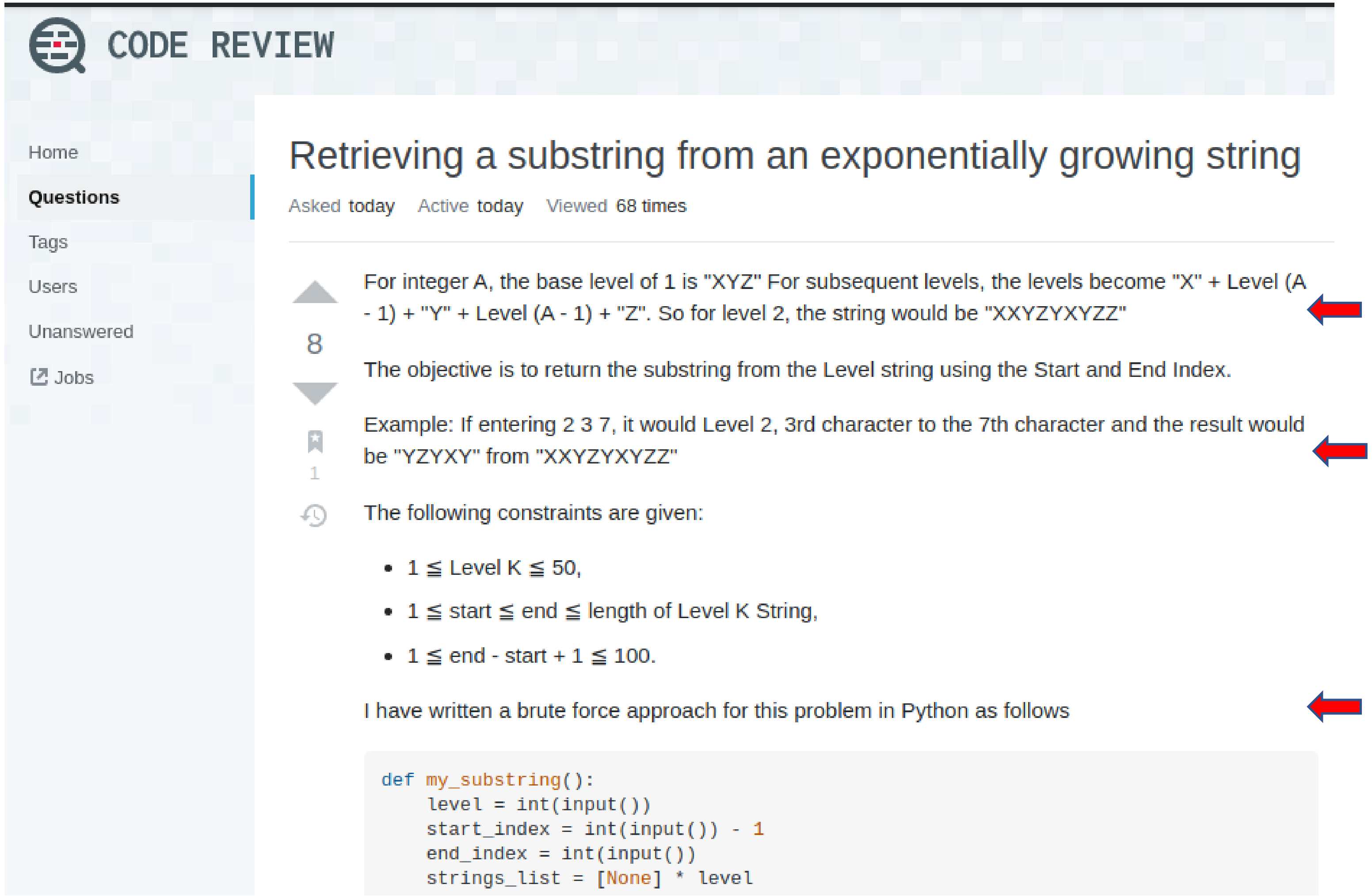
Worth Reading

SOTA is still not 100% solved

It's not a good time. (It's never a good time).

In a quiet voice, he said "this will not work. I am quitting", and then he left the room.

Real world example



Retrieving a substring from an exponentially growing string

Asked today Active today Viewed 68 times

For integer A, the base level of 1 is "XYZ" For subsequent levels, the levels become "X" + Level (A - 1) + "Y" + Level (A - 1) + "Z". So for level 2, the string would be "XXYZYXYZZ"

The objective is to return the substring from the Level string using the Start and End Index.

Example: If entering 2 3 7, it would Level 2, 3rd character to the 7th character and the result would be "YZYXY" from "XXYZYXYZZ"

The following constraints are given:

- 1 ≦ Level K ≦ 50,
- 1 ≦ start ≦ end ≦ length of Level K String,
- 1 ≦ end - start + 1 ≦ 100.

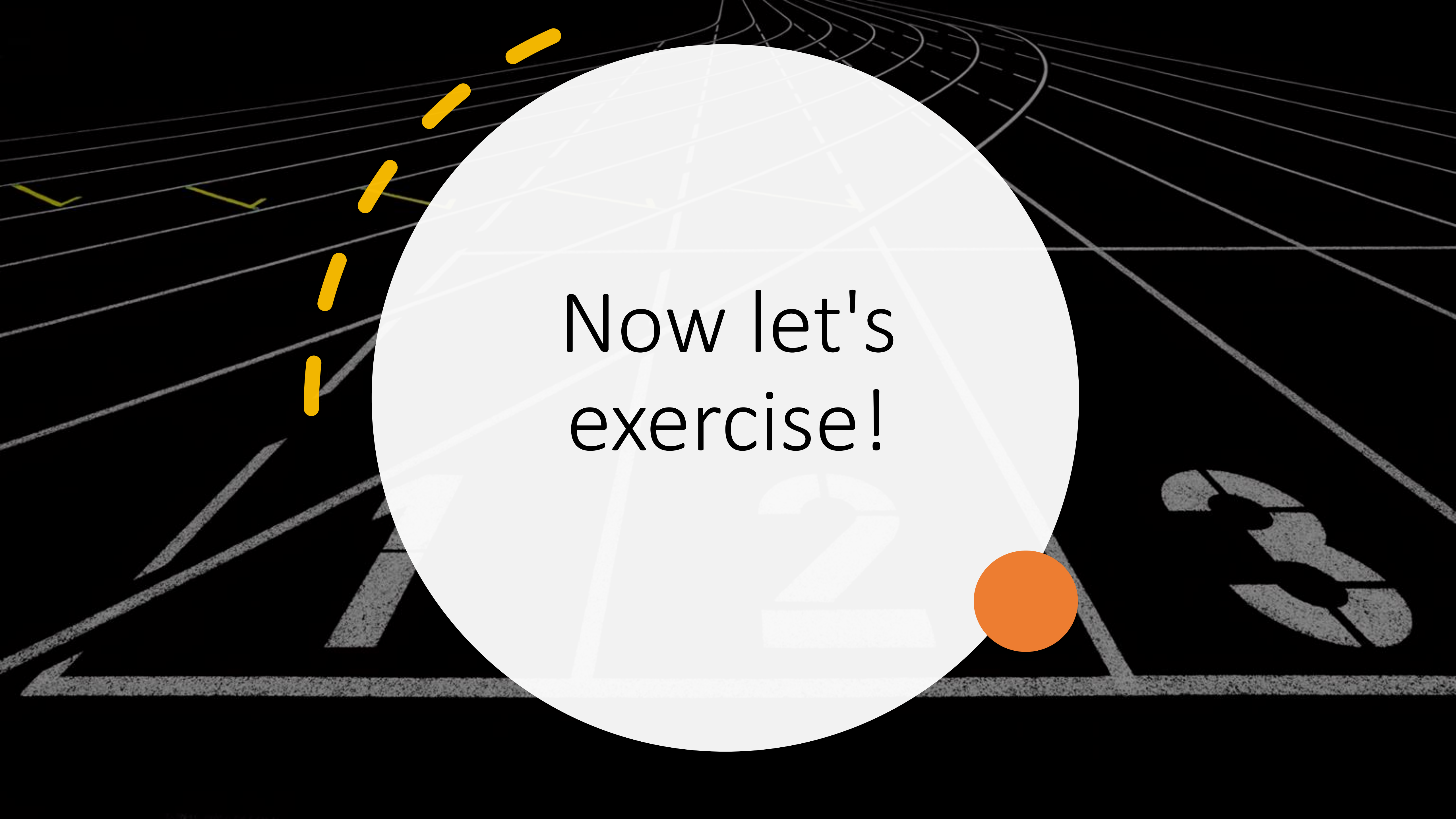
I have written a brute force approach for this problem in Python as follows

```
def my_substring():
    level = int(input())
    start_index = int(input()) - 1
    end_index = int(input())
    strings_list = [None] * level
```

Pop quiz

Name the 5 levels of Layers of Linguistic Knowledge
(and what do they mean)

- Phonology
- Morphology
- Syntax
- Semantics
- Pragmatics



Now let's exercise!