



# **Pendahuluan Pemrosesan Bahasa Alami**

Perkuliahan #1

Ade Romadhony

CII7G3 - Pemrosesan Bahasa Alami Lanjut  
Semester II 2020/2021



## **Kerangka Bahasan**

- Perkenalan Perkuliahan
- Pendahuluan Pemrosesan Bahasa Alami (PBA)



## **Mahasiswa yang mengikuti perkuliahan ini**

- Apakah sedang mengerjakan TA dengan topik terkait Pemrosesan Bahasa Alami?
- Apa yang membuat Anda tertarik mengambil mata kuliah ini?
- Apa yang Anda harapkan setelah mengikuti mata kuliah ini?



**Fakultas Informatika**  
School of Computing  
Telkom University

# **Pendahuluan PBA**



## Mesin Penerjemah / *Machine Translation*

The screenshot shows the Google Translate web interface. At the top left is the Google logo. To its right are a grid icon, a notification badge with the number '1', and a globe icon. Below the logo is the word 'Translate' in red, and to its right is a link that says 'Turn off instant translation'. The main interface features two language selection menus. The left menu has 'English', 'Indonesian', 'Spanish', and 'Detect language' with a dropdown arrow. The right menu has 'Indonesian', 'English', and 'Spanish' with a dropdown arrow. A blue 'Translate' button is positioned to the right of the second menu. Below the menus are two text boxes. The left box contains the Indonesian text 'dia datang ke sini untuk memberi tempe dan tahu' and has a close 'x' icon in the top right corner. Below this box are icons for voice input, speaker, keyboard, and a dropdown arrow. The right box contains the English translation 'he came here to give tempeh and tofu'. Below this box are icons for a star, copy, speaker, and share.

# Sistem tanya jawab / *Question Answering System (QAS)*



who is the ceo of telkom

[All](#) [News](#) [Images](#) [Maps](#) [Videos](#) [More ▾](#) [Search](#)

About 473,000 results (0.76 seconds)

Telkom Indonesia / CEO


**Alex J Sinaga**  
Dec 19, 2014—

who is the president of mit  

[All](#) [Images](#) [News](#) [Maps](#) [Videos](#) [More ▾](#) [Search tools](#)

About 89,100,000 results (0.54 seconds)

Massachusetts Institute of Technology / President

**L. Rafael Reif** 

About **President L. Rafael Reif**. Since July 2012, **Rafael Reif** has served as the 17th President of the Massachusetts Institute of Technology (MIT), where he is leading MIT's pioneering efforts to help shape the future of higher education.

[About President L. Rafael Reif | MIT Office of the President](#)  
[president.mit.edu/biography](http://president.mit.edu/biography)



## Sentiment Analysis

 **MylesCampbell** ❤️ @Thereal\_Cozycam · 56 min 

I really hate green iPhone messages

[Ver traducción](#)

---

RETWEETS 25    FAVORITOS 8



---

13:52 - 23 oct. 2015 · [Detalles](#)

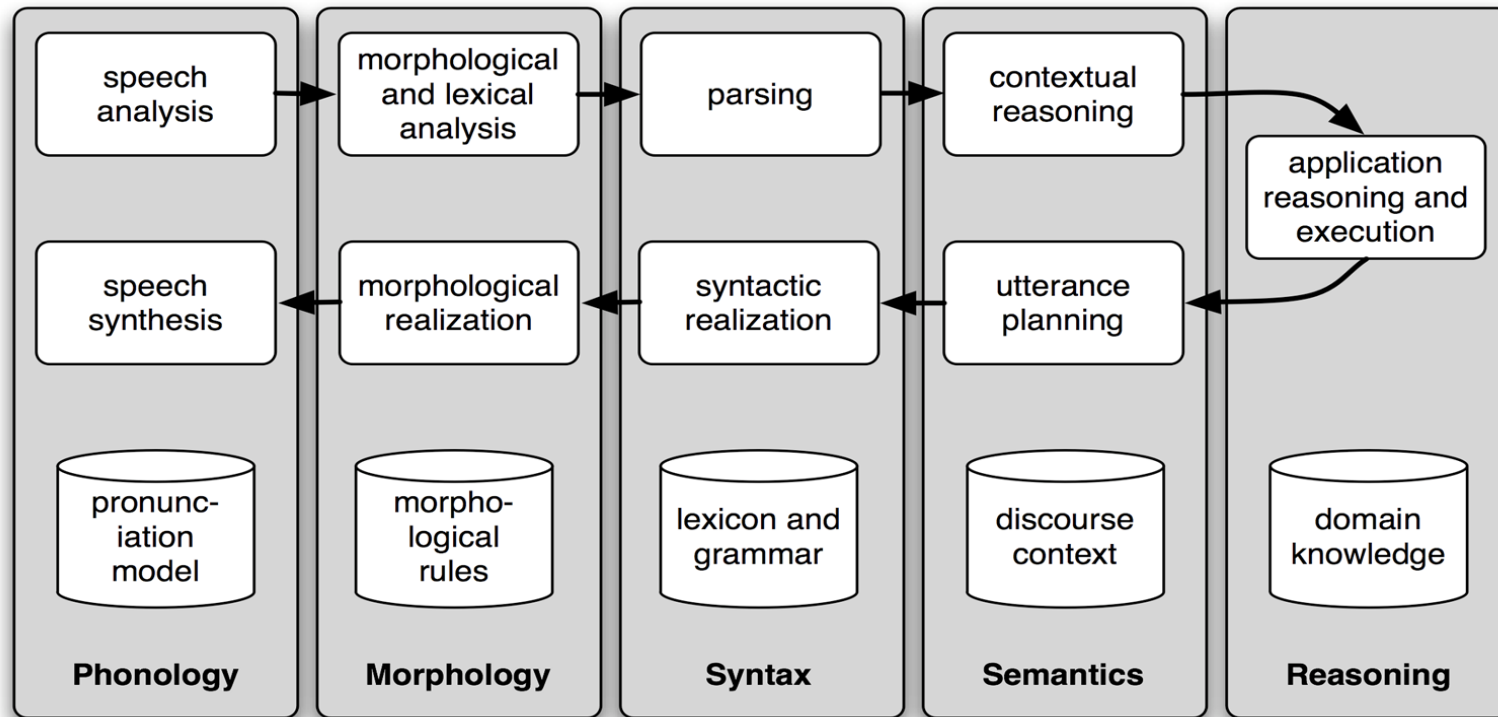
# Ultimate Dream: Conversational Agent JARVIS ?





# Ultimate Dream: Conversational Agent JARVIS ?

Source : Nltk book



## **Apa itu Pemrosesan Bahasa Alami (PBA)?**

Kumpulan metode untuk membuat bahasa manusia dapat “dimengerti” oleh komputer

Contoh-contoh aplikasi dalam kehidupan sehari-hari yang kita bahas sebelum ini menunjukkan bahwa PBA sudah menjadi bagian yang penting dalam kehidupan manusia saat ini.

## Beberapa bidang ilmu yang terkait dengan PBA (1)

- Linguistik komputasional
  - Apakah topik yang sama dengan PBA tetapi beda istilah saja?
- Pembelajaran Mesin
  - Pendekatan/metode yang digunakan untuk pekerjaan PBA masa kini sangat bergantung pada perkembangan pembelajaran mesin
- Sains Komputer
  - Ingat saat belajar TBA? *Finite State* dan *Pushdown Automata* merupakan dasar untuk banyak aplikasi PBA

## Beberapa bidang ilmu yang terkait dengan PBA (2)

- Pemrosesan Suara
  - Sinyal suara perlu diubah dulu ke teks untuk diproses lebih lanjut. Pemrosesan lebih lanjut -> pemrosesan teks bahasa (PBA)
- Penambangan Teks
  - Umumnya istilah ini dipakai untuk penerapan teknik penambangan data pada teks. Perbedaan dengan PBA, penambangan teks umumnya lebih memperhatikan kecepatan dan skalabilitas algoritma, dibanding struktur tata bahasa
- Sistem Rekomendasi
  - Sistem rekomendasi pada dataset berupa teks, mirip dengan keterkaitan teks
- Analisis jejaring sosial
  - Pesan teks yang dituliskan oleh pengguna termasuk salah satu sumber penting dalam analisis

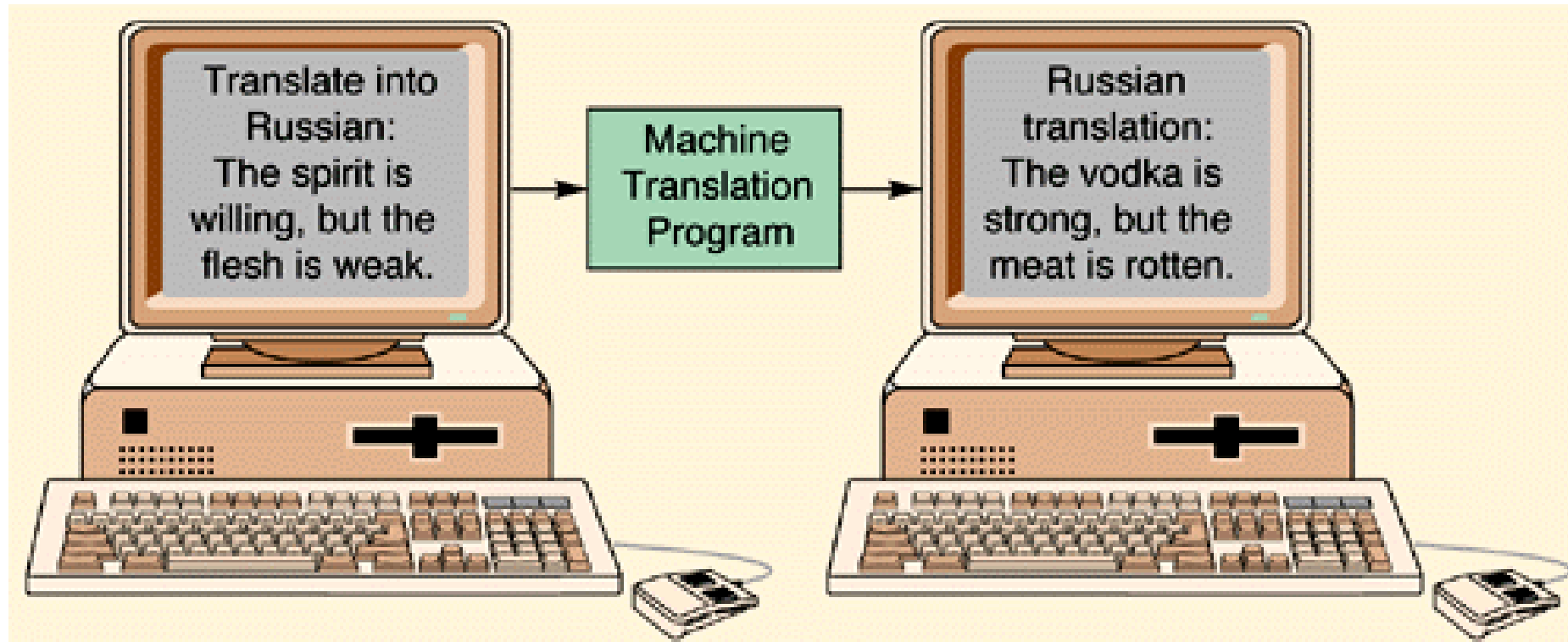


## **Istilah-istilah dalam PBA**

- *Phonetics and Phonology – The study of linguistic sounds.*
- *Morphology – The study of the meaningful components of words.*
- *Syntax – The study of the structural relationships between words.*
- *Semantics – The study of meaning.*
- *Pragmatics – The study of how language is used to accomplish goals.*
- *Discourse – The study of linguistic units larger than a single utterance*



## Contoh persoalan PBA: mesin penerjemah (1)



## Contoh persoalan PBA: ekstraksi informasi(2)

who is the dean of telkom university school of computing



[All](#) [News](#) [Images](#) [Maps](#) [Videos](#) [More](#) [Settings](#) [Tools](#)

About 101,000 results (0.42 seconds)

[io.telkomuniversity.ac.id](http://io.telkomuniversity.ac.id) > [faculty-of-computing](#) ▾

### School of Computing - Telkom University International Office

**School of Computing** Informatics in 2014, has three courses namely S1 Computer Science, S1 and S2 computing Science Informatics. Being a world-class ...

Missing: [dean](#) | Must include: [dean](#)




Fakultas Informatika Universitas Telkom  
[soc.telkomuniversity.ac.id](http://soc.telkomuniversity.ac.id)

## Contoh persoalan PBA: *speech recognition* (3)

Voice Model: US English broadband model(16KHz) ▼      Keywords to spot: sense of pride, watson, technology, changing the world, round, w

Detect multiple speakers (Not supported on current model)

Record Audio   Upload Audio File   Play Sample 1   Play Sample 2

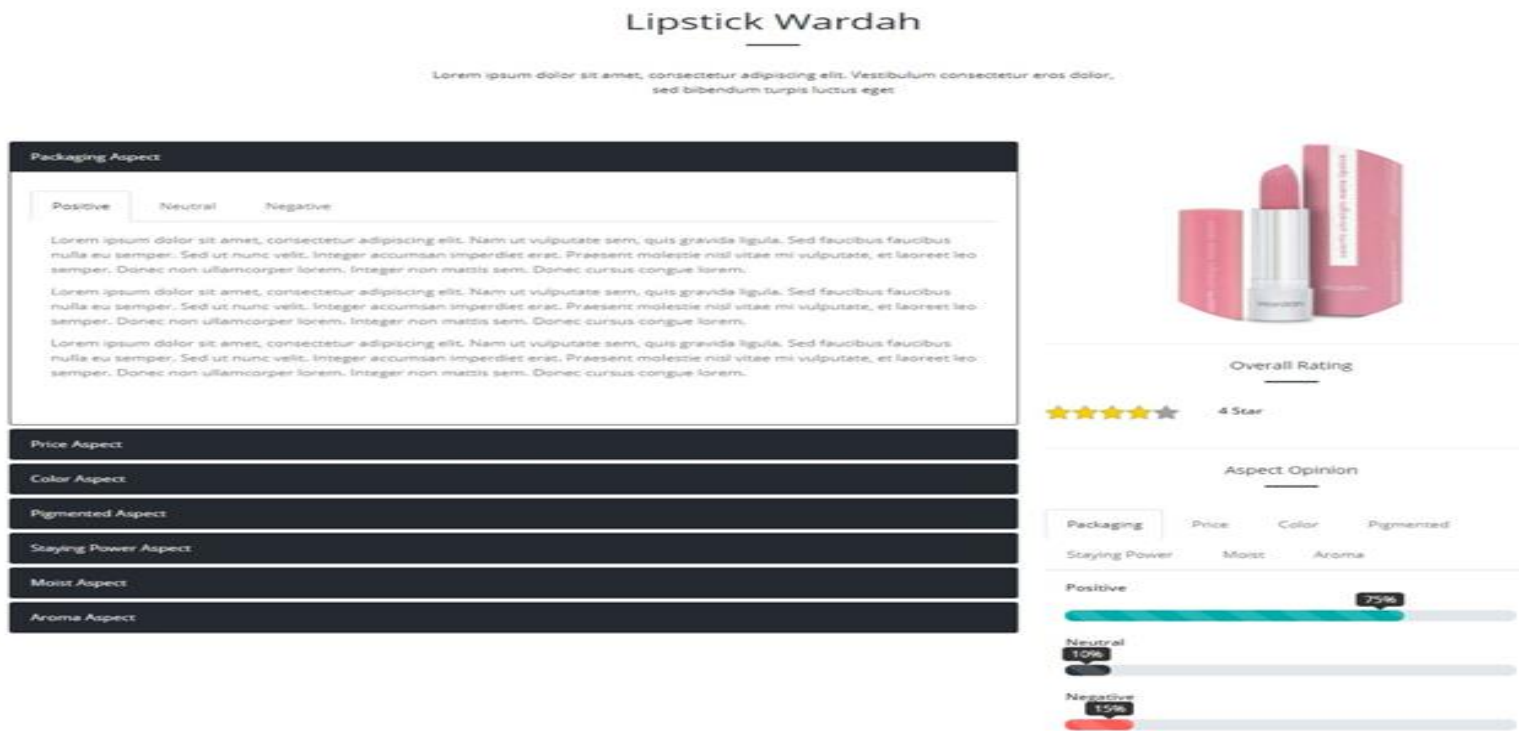
 No speech detected for 30s.

Text	Word Timings and Alternatives	Keywords (0/7)	JSON
I love you. You love mean. Where had bees and lo Li. We take great BP had. And the keys from me do you. Want you say you Love Me E. to. I love you you Love Me we are best friends like friends who'd pee wee that Greg being. And the keys from me do you want you say you Love Me too.			



## Contoh penelitian/proyek yang pernah dikerjakan

- Ringkasan ulasan produk berbasis aspek





## Contoh perkembangan terkini: GPT-3

- *Generative Pre-trained Transformer 3 is an autoregressive language model that uses deep learning to **produce human-like text**.* [dikembangkan oleh OpenAI]
- *Task yang dikerjakan?* -> pada dasarnya adalah melengkapi teks
- *Dapat menghasilkan teks apa saja?*
  - Cerita
  - Lagu
  - Tulisan teknis
  - Kode program
  - dll



## Mengapa terdapat persoalan-persoalan pada PBA?

### **Ambiguitas ditemui di semua level !!!**

#### *Phonetics and Phonology*

- I Scream vs Ice cream

#### *Morphology*

- Unionized = union + ized vs un+ionized

#### *Syntax*

- Squad helps [dog bite victim] vs [Squad helps dog] bite victim

#### *Semantics*

- Jack invited Mary to the Halloween ball



## Mengapa terdapat persoalan-persoalan pada PBA?

### **Ambiguitas ditemui di semua level !!!**

#### *Discourse*

- Merck & Co. formed a joint venture with Ache Group, of Brazil. **It** will be called Prodome Ltd

#### *Pragmatics*

- Bagaimana kalimat digunakan di situasi yang berbeda  
“I just came from New York”
  - Would you like to go to New York today?
  - Would you like to go to Boston today?
  - Why do you seem so out of it?
  - Boy, you look tired.

# Klasifikasi teknologi bahasa (dari slide Jurafsky)

## mostly solved

**Spam detection**

Let's go to Agra! 

Buy VIAGRA ... 

**Part-of-speech (POS) tagging**

ADJ ADJ NOUN VERB ADV

Colorless green ideas sleep furiously.


**Named entity recognition (NER)**


PERSON ORG LOC

Einstein met with UN officials in Princeton

## making good progress

**Sentiment analysis**


Best roast chicken in San Francisco! 

The waiter ignored us for 20 minutes. 

**Coreference resolution**

Carter told Mubarak he shouldn't run again.

**Word sense disambiguation**

I need new batteries for my *mouse*. 

**Parsing**


I can see Alcatraz from the window!

**Machine translation (MT)**

第13届上海国际电影节开幕... 

The 13<sup>th</sup> Shanghai International Film Festival...

**Information extraction (IE)**

You're invited to our dinner party, Friday May 27 at 8:30  Party May 27 [add](#)

## still really hard

**Question answering (QA)**

Q. How effective is ibuprofen in reducing fever in patients with acute febrile illness?

**Paraphrase**

XYZ acquired ABC yesterday

ABC has been taken over by XYZ

**Summarization**

The Dow Jones is up


The S&P500 jumped

Housing prices rose

⇒ Economy is good

**Dialog**

Where is Citizen Kane playing in SF?

Castro Theatre at 7:30. Do you want a ticket? 



## Sebab lain mengapa pemahaman Bahasa sulit? (dari slide Jurafsky)

### non-standard English

Great job @justinbieber! Were SOO PROUD of what youve accomplished! U taught us 2 #neversaynever & you yourself should never give up either♥

### segmentation issues

the New York-New Haven Railroad  
the New York-New Haven Railroad

### idioms

dark horse  
get cold feet  
lose face  
throw in the towel

### neologisms

unfriend  
Retweet  
bromance

### world knowledge

Mary and Sue are sisters.  
Mary and Sue are mothers.

### tricky entity names

Where is *A Bug's Life* playing ...  
*Let It Be* was recorded ...  
... a mutation on the *for* gene ...

But that's what makes it fun!



## Perkembangan dalam PBA

- *Task*-nya sulit untuk diselesaikan! Apa saja yang diperlukan?
  - Pengetahuan tentang bahasa
  - Sumber pengetahuan/*knowledge resources* yang dapat dimanfaatkan
  - Bagaimana untuk mengkombinasikan sumber pengetahuan
- Secara umum, bagaimana menyelesaikan *task* PBA:
  - Model probabilistik yang dibangun dari data:
    - $P(\text{"maison"} \rightarrow \text{"house"})$  **high**
    - $P(\text{"L'avocat général"} \rightarrow \text{"the general avocado"})$  **low**
  - Untungnya, umumnya fitur leksikal saja dapat menyelesaikan setengah pekerjaan untuk menyelesaikan *task*



## **Apa isi mata kuliah ini secara umum?**

- Pengantar ke topik Pemrosesan Bahasa Alami
- Mempelajari teknik untuk mengatasi persoalan utama dalam PBA: ambiguitas



## Diskusi: PBA pada bahasa selain Bahasa Inggris

- Bahasan di buku referensi atau artikel-artikel ilmiah, umumnya mengacu pada Bahasa Inggris
- Bagaimana dengan PBA untuk Bahasa selain Bahasa Inggris?
  - Apakah jika kinerja sebuah sistem pemrosesan Bahasa Inggris sudah sangat baik (misal akurasi > 90%), maka otomatis jika digunakan metode yang sama pada bahasa selain Bahasa Inggris, hasilnya akan lebih baik?

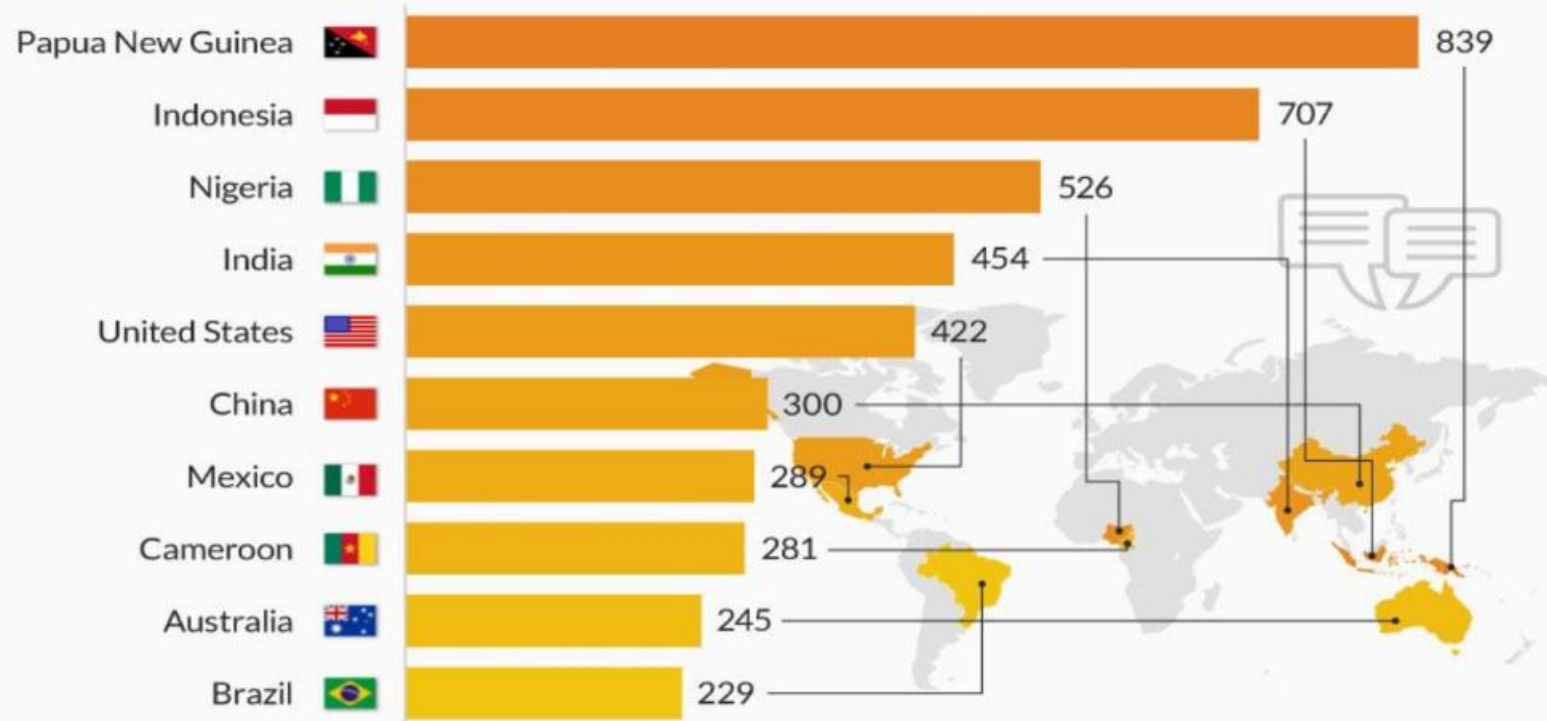
Bahan bacaan diskusi:

*Why you should do NLP Beyond English* <https://runder.io/nlp-beyond-english/>

# Negara-negara dengan bahasa daerah paling banyak

## The Countries With The Most Spoken Languages

Number of living languages spoken per country in 2015



sumber: ethnologue



## Diskusi: PBA pada bahasa selain Bahasa Inggris

- Bahasa daerah apa saja yang ada di Indonesia?
  - Jawa
  - Sunda
  - Bali
  - Batak: sekitar 5 varian Bahasa
  - Makassar: Bugis dan Makassar
  - .....

## Tugas Membaca

- Bukalah website Buku Speech and Language Processing 3<sup>rd</sup> edition <https://web.stanford.edu/~jurafsky/slp3/>
- Pilih salah satu chapter yang sesuai dengan materi yang akan kita pelajari selama satu semester ini
- Presentasikan dengan singkat:
  - Definisi/penjelasan *task* tersebut. Misal: POSTagging adalah ....
  - Contoh *input* dan *output*.
  - Tantangan penyelesaian *task* (sulitnya di mana?)
  - Metode yang dapat digunakan untuk menyelesaikan *task* tersebut