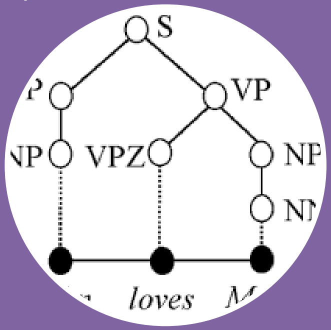
Bootstrapping Asian Language Treebank using Indonesian language resource

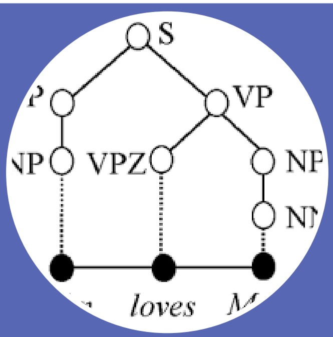
Badan Pengkajian dan Penerapan Teknologi

Hammam Riza
Deputy Chairman
Agency for the Assessment and Application of Technology (BPPT)

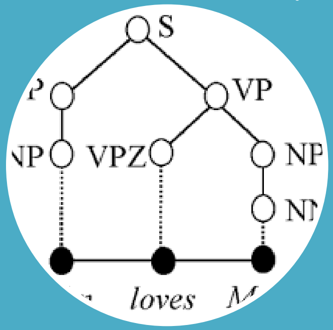
Treebank



Treebanks have become valuable resources in natural language processing (NLP) in recent years (Abeillé, 2003) such as training corpus, repository for lingusitic research or as evaluation corpus.



A treebank is a collection of syntactically annotated sentences in which the annotation has been manually checked. The name derives from the fact that syntactic descriptions of sentences often come in the form of tree structures, in particular constituent trees.



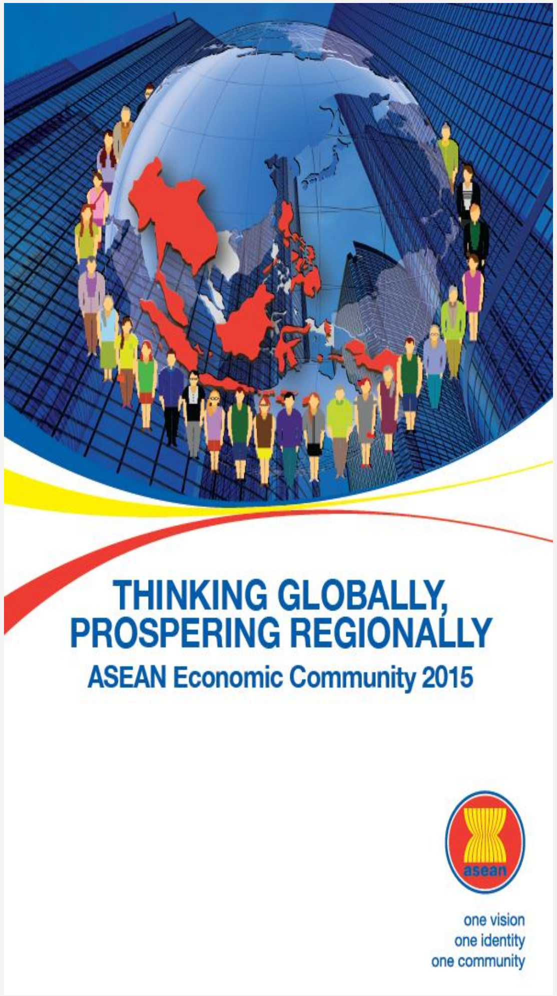
But treebank annotation has also been done in the framework of dependency grammar and recent annotation has also exceeded syntax towards semantic features such as predicate-argument structures or word senses.

Introduction

- **READY:** This is ASEAN’s time. In the geographic heart of the world’s premier growth corridor, ASEAN is poised to “seize the moment.”

- **SET:** With a market of **over 600 million consumers** and combined **GDP of nearly US$3 trillion**, ASEAN is offering a future of prosperity and stability. The AEC is one of the foundations of that future.

- **GO:** Agreements on trade, services and investment are changing the economic landscape and allowing the freer flow of goods, services and people across the region.



THINKING GLOBALLY,
PROSPERING REGIONALLY
ASEAN Economic Community 2015

asean

one vision
one identity
one community

The ASEAN Economic Community (AEC) shall be the goal of regional economic integration by 31 Dec 2015. AEC envisages the following key characteristics:

(a) a single market and production base,

(b) a highly competitive economic region,

(c) a region of equitable economic development,

(d) a region fully integrated into the global economy.

Total Population: 600 million+
GDP: USD $3 trillion

Extending ASEAN+ICJK = Asian Languages

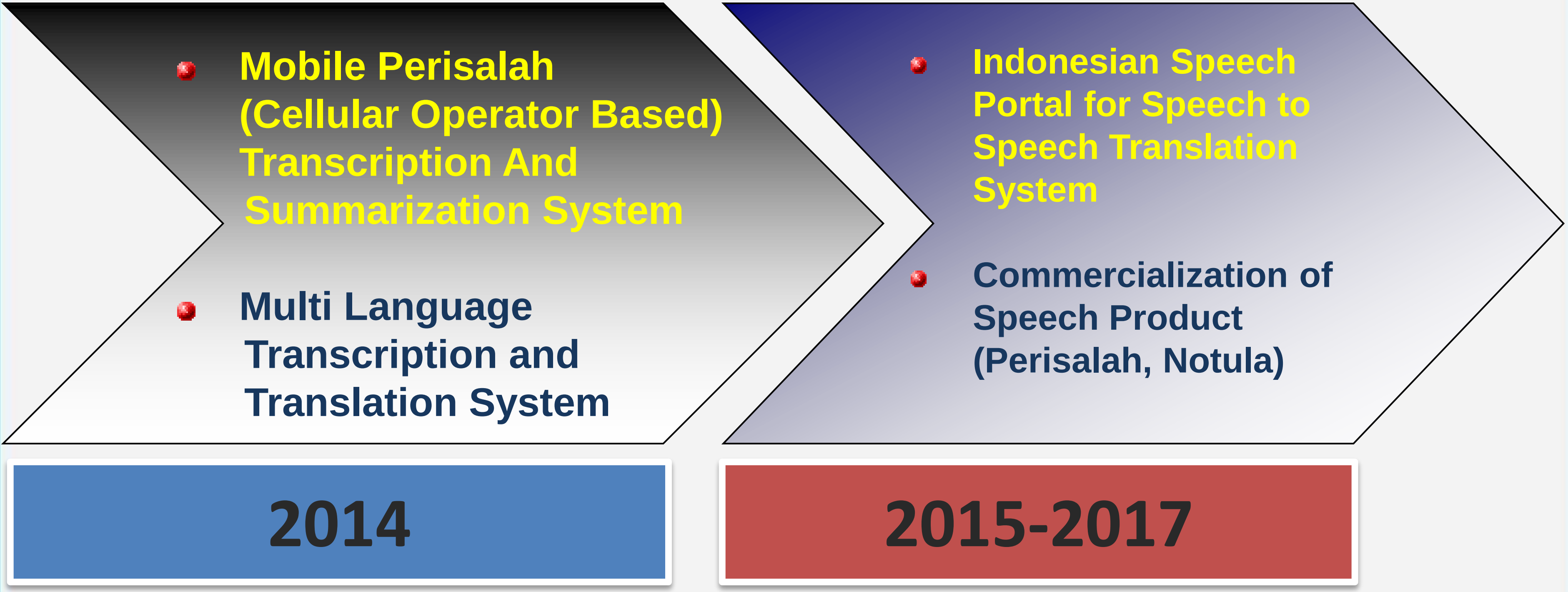
Why we need Asian Language Treebank (ALT)?

- **Accelerates research of NLP** for Asian languages
  - Indonesian, Vietnamese, Japanese, Khmer, Laos, Malay, Myammar, Philippine, Thai, ….
- **No** publicly available POS-tagged and constituency **tree corpora** for most of Asian languages. (Though, some corpora are available for some languages)
- **No parallel corpora** among all Asian languages
- Expected members
  - NICT, BPPT, IOIT, NECTEC, UCSY, and other research bodies
  - NICT and UCSY have already started making parts of ALT
  - NICT and IOIT agreed to propose the ALT project to ASEAN IVO

Schedule for ALT

- Current progress in FY 2015
  - NICT translates English Wikinews (460,000 words, 20,000 sentences) into Indonesian, Vietnamese , Japanese, Thai, Khmer, Laos, Malay, Myammar, Philippine,
  - NICT makes the Japanese and English treebanks
  - UCSY makes the Myammar treebank
- Proposal for project in FY 2016
  - NICT provides Indonesian-English Wikinews parallel corpus to BPPT
  - NICT provides Vietnamese-English Wikinews parallel corpus to IOIT
  - BPPT makes the Indonesian language treebank
  - IOIT makes the Vietnamese language treebank
    - Word segmentaion, POS tagging, Parsing, Word alignment (NICT has annotation tools)
  - NICT develops Japanese and English NLP tools using Japanese and English treebanks
  - ALT and tools are shared with NICT, BPPT, IOIT and other participants

Speech Technology R&D Roadmap



- Mobile Perisalah (Cellular Operator Based) Transcription And Summarization System
- Multi Language Transcription and Translation System

2014

- Indonesian Speech Portal for Speech to Speech Translation System
- Commercialization of Speech Product (Perisalah, Notula)

2015-2017

Universal Speech Translation Advance Research (U-STAR)

Speech Corpora, Parallel Text Corpora ,TTS, ASR

Speech Recognition and NLP (continuing activities)

- **Speech recognition.**

Leveraging our years of experience with Indonesian language models, we are currently developing acoustic models trained on a largemulti-speaker speech corpus, and investigating thesuitability of applying these models to existing open-source speech recognition systems such as JULIUS3 and SPHINX4

- **Corpus-based NLP tools for Information Retrieval.**

In a joint collaboration with NUS and USM, vari-ous resources and algorithms are being researched forlarge-scale Malay and Indonesian information retrievalusing corpus-based methods. Interim results includethe development of a statistical part-of-speech tagger.

- **Construction of an Indonesian WordNet.**

An ongoing project is concerned with the development of an Indonesian WordNet. Using the expand model approach], map Princeton WordNet to existing word sense definitions in the Indonesian Master Dictionary (KBBI) ,which defines semantic equivalence classes between KBBI senses.

- **Finite state morphological analysis.**

Reduplication is an example of non-concatenative morphology, which formally cannot be modelled by finite state techniques, the prevalent paradigm for modern morphological analysis.

Indonesian Language Tools

- **Language Processing Tools**
  - Indonesian Stemmer
  - Indonesian POS Tagger
  - Indonesian Named Entity Tagger for text article
  - Indonesian Phrase Chunker
  - Indonesian Statistical Constituent Parser
  - Indonesian Dependency Parser
  - Indonesian Reference Resolution
  - Indonesian Semantic Analysis (Constituent based)
  - Indonesian NE Tagger for social media
  - Indonesian Term Normalization for social media
  - Indonesian Spelling Checker
- **Text Mining**
  - Indonesian Question Answering System (open domain & closed domain of dialogue system)
  - Indonesian Information Extraction (Social Media): Citizen Complaint, Strike Event
  - Indonesian Text Classification (Social Media)

- **Speech**
  - Indonesian Automatic Speech Recognizer using Deep Neural Network
  - Indonesian ASR for spontaneous speech
  - Indonesian Speech Synthesizer using Hidden Markov Model
  - Improvement of Indonesian Prosody using Decision Tree Learning
  - Al-Quran Automatic Speech Recognizer: corpus and system development
  - Indonesian emotional recognition: corpus and system development

ITB Indonesian TreeBank

- Consists of Lemma, POS Tag and Dependency Relation Information
- 2098 sentences, annotated by 5 native speakers which are Indonesia linguists

Sentence Type		Number of Sentences (& Percentage)
Sentence type	Single	1067 (50,86%)
	Compound – equivalent	349 (16,63%)
	Compound – gradual	527 (25.12%)
	Compound – complex	155 (7,39%)
Type of Sentence Root	Transitive Verb	1017 (48,47%)
	Intransitive Verb	989 (47,14%)
	Adjektive	69 (3,29%)
	Others	23 (1,10%)

Information of Data Tuple

- ID: token position in sentence
- FORM: token lexical
- LEMMA: base form of token
- CPOSTAG: coarse-grained POS tag (universal POS tag universal)
- POSTAG: fine-grained POS tag
- FEATS: syntactic information (still empty)
- HEAD: head ID of current token
- DEPREL: dependency relation type (still empty)
- PHEAD: head ID if it is projective
- PDEPREL: dependency relation type of PHDEAD

Example of Data Tuple

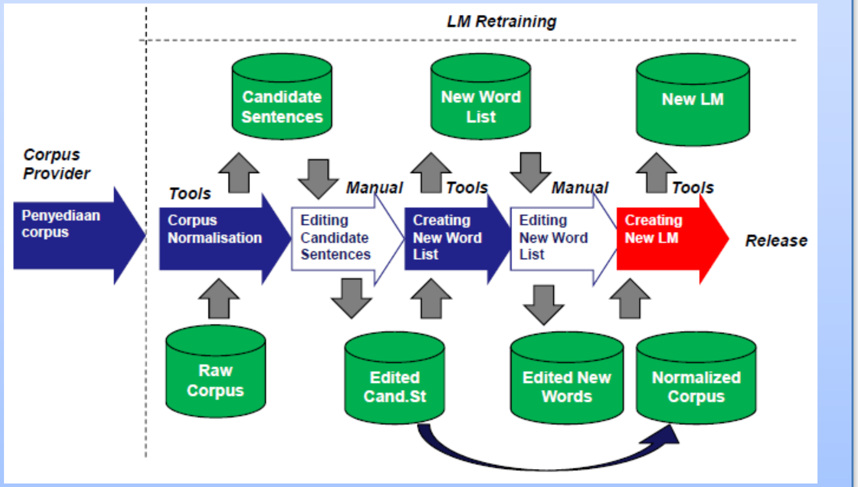
```
1 Adik Adik NOUN NNP _ 2 _ _ _
2 membeli beli VERB VBT _ 0 _ _ _
3 susu susu NOUN NN _ 2 _ _ _
4 kemarin kemarin NOUN NN _ 2 _ _ _

1 Jangan Jangan ADV NEG _ 2 _ _ _
2 bermain-main main VERB VBI _ 0 _ _ _
3 saja saja ADV RB _ 2 _ _ _
4 kamu kamu PRON PRP _ 2 _ _ _

1 Ketika Ketika ADP SC _ 3 _ _ _
2 saya saya PRON PRP _ 3 _ _ _
3 pulang pulang VERB VBI _ 7 _ _ _
4 ke ke PRT IN _ 3 _ _ _
5 rumah rumah NOUN NN _ 4 _ _ _
6 hari hari NOUN NN _ 7 _ _ _
7 hujan hujan NOUN NN _ 0 _ _ _
```

Corpus Development 2013-2015

- Perisalah Speech Corpora
- Perisalah POS Tagged Corpus
- Corpus Management System

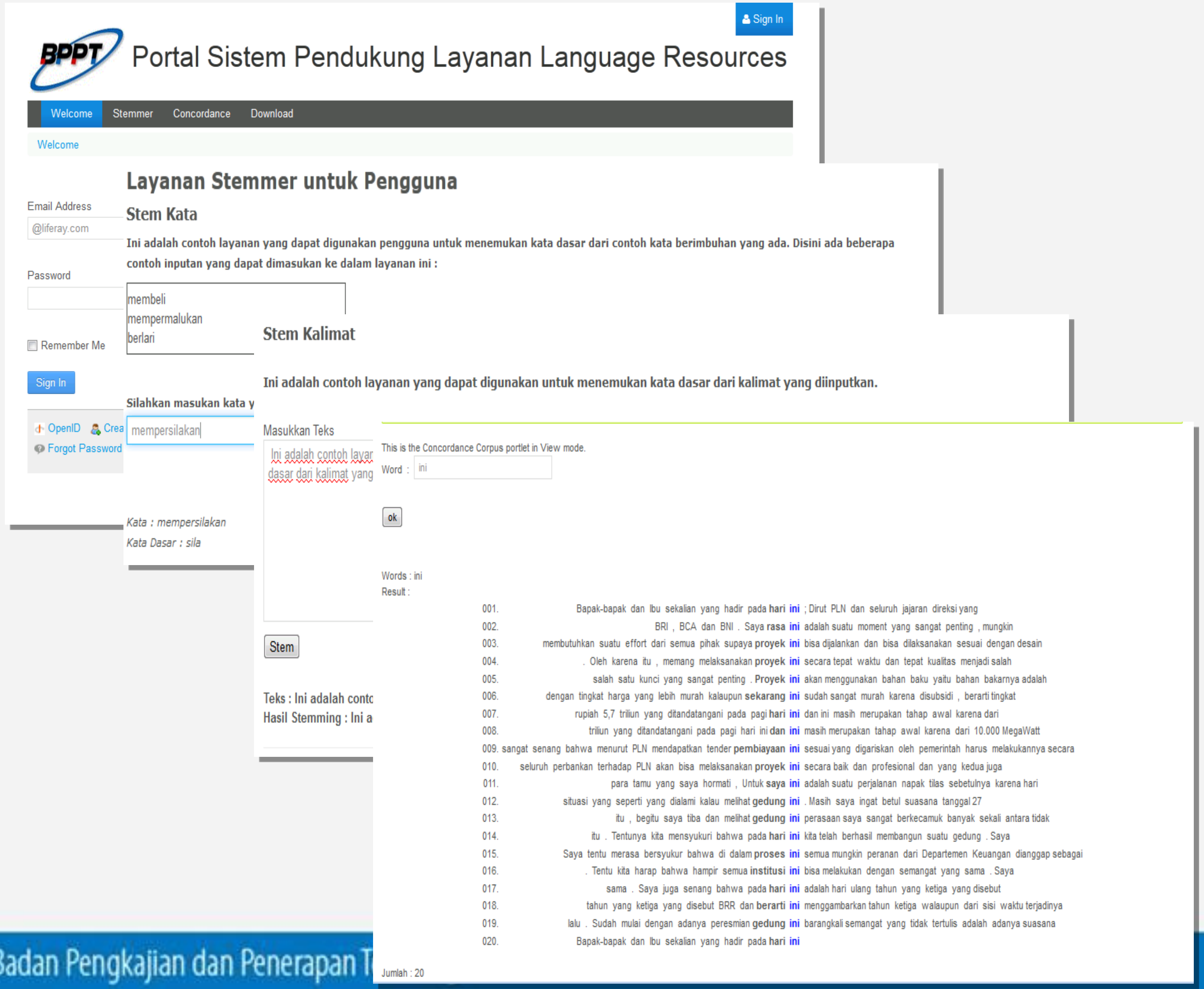


- Indonesian POS Tagged Corpus
- Indonesian Named Entity Tagged Corpus
- Indonesian Syntactical Tree Tagged Corpus
- Indonesian Dependency Tree Tagged Corpus
- Indonesian Question Answering
- Using of Indonesian TTS for blind operator who work in Call Center
- Sunda Lexical Database
- Indonesian-Japanese Parallel Corpus
- Indonesian-English Parallel Corpus

INACL – Community Setup

- Background:
  - Importance on research collaboration among Indonesian CL researchers, government and industry
- Activities:
  - Initiated (first meeting) at PACLING 2015 (Bali, organized by BPPT & ITB)
  - Communication group: social media & mailing list
  - National workshop meeting (plan: December 2015 or January 2016)
  - Indonesian Data Resource & CL Tools Sharing
  - Support international conference (such as Cocosda 2016)
- Website:
  - http://www.inacl.id
- Member:
  - 91 members (from 32 universities + 1 government inst. + 5 industries)

Portal for Language Resource Service Support System



Portal Sistem Pendukung Layanan Language Resources

Email Address: @liferay.com

Password:

□ Remember Me

Sign In

OpenID Crea

Forgot Password

Layanan Stemmer untuk Pengguna

Stem Kata

Ini adalah contoh layanan yang dapat digunakan pengguna untuk menemukan kata dasar dari contoh kata berimbuhan yang ada. Disini ada beberapa contoh inputan yang dapat dimasukan ke dalam layanan ini :

membeli
mempermalukan
berlari

Silahkan masukan kata y: mempersilakan

Kata : mempersilakan
Kata Dasar : sila

Stem Kalimat

Ini adalah contoh layanan yang dapat digunakan untuk menemukan kata dasar dari kalimat yang diinputkan.

Masukkan Teks: Ini adalah contoh layar dasar dari kalimat yang

Stem

Teks : Ini adalah conto
Hasil Stemming : Ini a

This is the Concordance Corpus portlet in View mode.

Word : ini

ok

Words : ini
Result :

```
001.                Bapak-bapak dan Ibu sekalian yang hadir pada hari ini ; Dirut PLN dan seluruh jajaran direksi yang
002.                                   BRI , BCA dan BNI . Saya rasa ini adalah suatu moment yang sangat penting , mungkin
003.        membutuhkan suatu effort dari semua pihak supaya proyek ini bisa dijalankan dan bisa dilaksanakan sesuai dengan desain
004.                 . Oleh karena itu , memang melaksanakan proyek ini secara tepat waktu dan tepat kualitas menjadi salah
005.                  salah satu kunci yang sangat penting . Proyek ini akan menggunakan bahan baku yaitu bahan bakarnya adalah
006.       dengan tingkat harga yang lebih murah kalaupun sekarang ini sudah sangat murah karena disubsidi , berarti tingkat
007.             rupiah 5,7 triliun yang ditandatangani pada pagi hari ini dan ini masih merupakan tahap awal karena dari
008.                triliun yang ditandatangani pada pagi hari ini dan ini masih merupakan tahap awal karena dari 10.000 MegaWatt
009. sangat senang bahwa menurut PLN mendapatkan tender pembiayaan ini sesuai yang digariskan oleh pemerintah harus melakukannya secara
010.    seluruh perbankan terhadap PLN akan bisa melaksanakan proyek ini secara baik dan profesional dan yang kedua juga
011.                       para tamu yang saya hormati , Untuk saya ini adalah suatu perjalanan napak tilas sebetulnya karena hari
012.            situasi yang seperti yang dialami kalau melihat gedung ini . Masih saya ingat betul suasana tanggal 27
013.                        itu , begitu saya tiba dan melihat gedung ini perasaan saya sangat berkecamuk banyak sekali antara tidak
014.                 itu . Tentunya kita mensyukuri bahwa pada hari ini kita telah berhasil membangun suatu gedung . Saya
015.            Saya tentu merasa bersyukur bahwa di dalam proses ini semua mungkin peranan dari Departemen Keuangan dianggap sebagai
016.                 . Tentu kita harap bahwa hampir semua institusi ini bisa melakukan dengan semangat yang sama . Saya
017.                      sama . Saya juga senang bahwa pada hari ini adalah hari ulang tahun yang ketiga yang disebut
018.                 tahun yang ketiga yang disebut BRR dan berarti ini menggambarkan tahun ketiga walaupun dari sisi waktu terjadinya
019.              lalu . Sudah mulai dengan adanya peresmian gedung ini barangkali semangat yang tidak tertulis adalah adanya suasana
020.                Bapak-bapak dan Ibu sekalian yang hadir pada hari ini
```

Jumlah : 20

Fitures :

- Download Corpus
  - Speech Corpus (Male-Female)
  - Monolingual
  - Bilingual
- Stemmer
- Concordance

Harigato Gozaimasu

THANK YOU