02. NLP: LEVEL OF LANGUAGE

FIRDAUS SOLIHIN

UNIVERSITAS TRUNOJOYO MADURA

LINGUISTICS

- Linguistik adalah studi ilmiah tentang bahasa manusia.
- Linguistik dapat secara luas dipecah menjadi tiga kategori atau subbidang studi:
  - Bentuk bahasa (*Language form*)
  - Makna bahasa (*Language meaning*)
  - Bahasa dalam konteks (*Language on context*)

What actually linguistics is?

Sound	Structure	Meaning
Phonetics	Morphology	Semantics
Phonology	Syntax	Pragmatics

PEMBENTUK BAHASA

- ABJAD (a,i,u,e,o,b,c,d,f)
- KATA (saya, belajar)
- KALIMAT (saya sedang belajar NLP)
- PARAGRAF (gabungan beberapa kalimat)
- BAHASA

Aplikasi NLP

- Text-based application
- Speech-based application

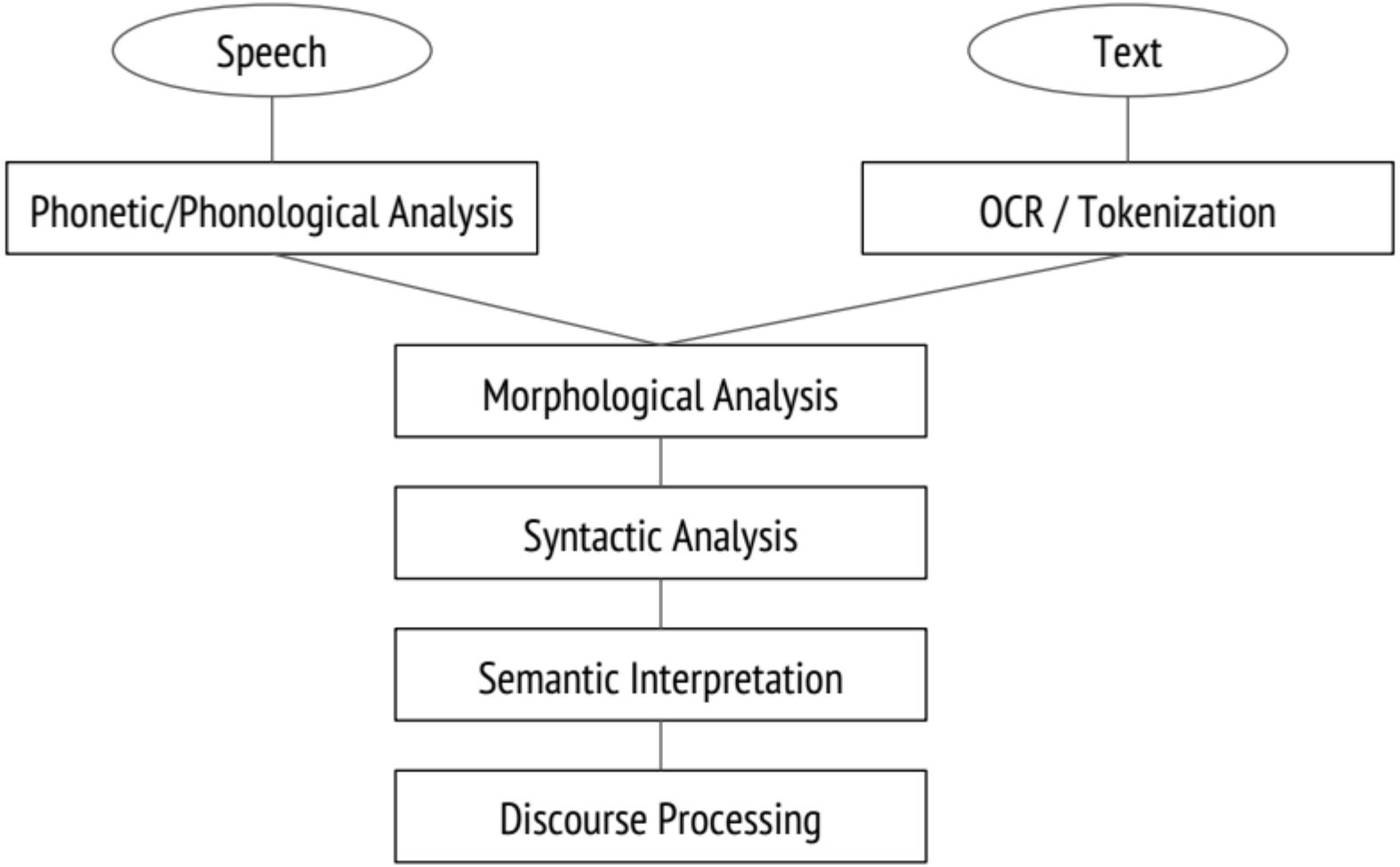
Text-based application

- Aplikasi yang melakukan memprosesan terhadap teks tertulis
- Contoh:
  - Mencari topik tertentu dari buku di perpustakaan
  - Mencari isi dari suatu berita atau artikel
  - Mencari isi dari email
  - Menterjemahkan dokumen dari suatu bahasa ke bahasa lain

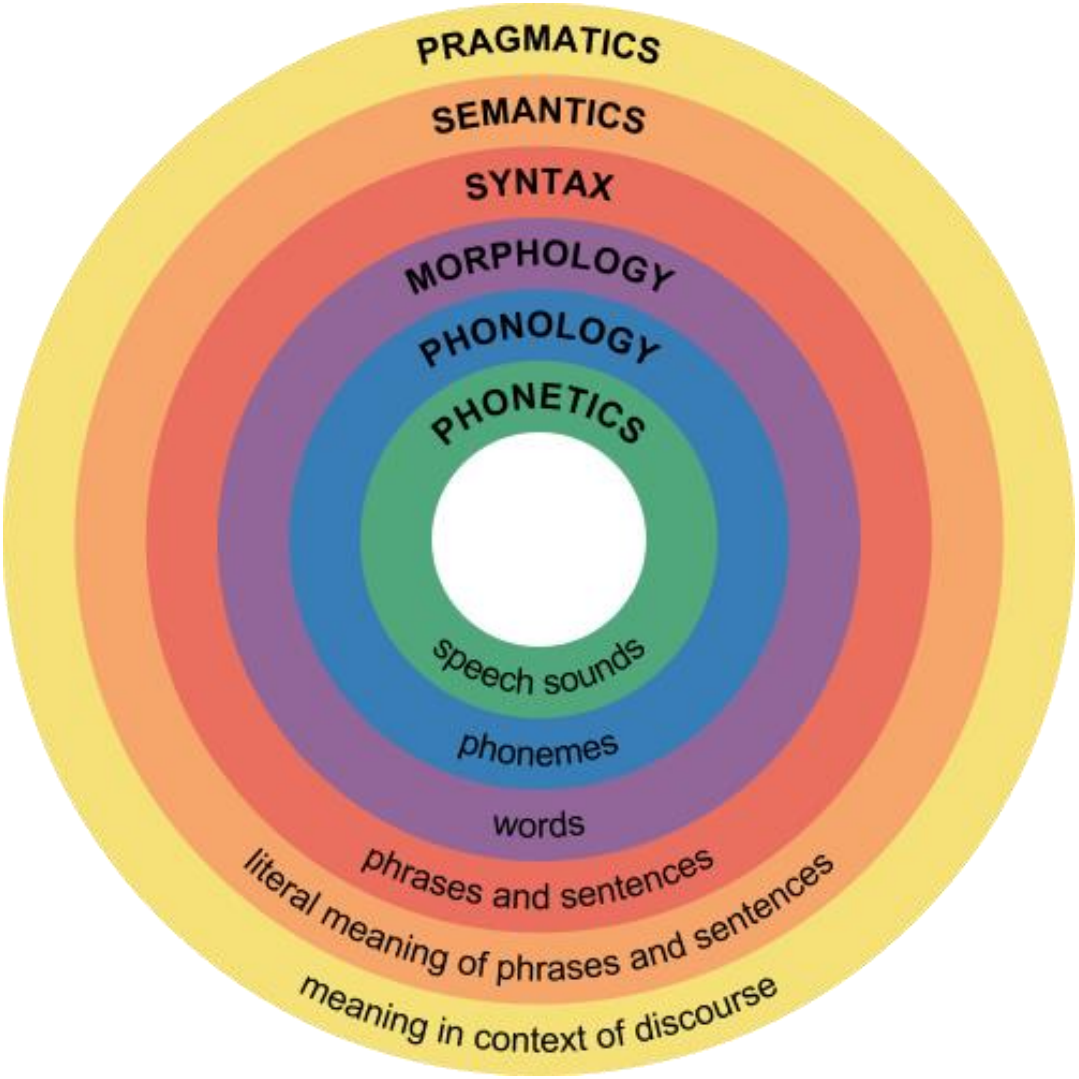
Speech-based application

- Aplikasi yang melakukan memprosesan dari bahasa lisan atau pengenalan suara.
- Contoh:
  - Sistem otomatis pelayanan melalui telepon
  - Control suara pada peralatan elektronik -Aplikasi peningkatan kemampuan berbahasa

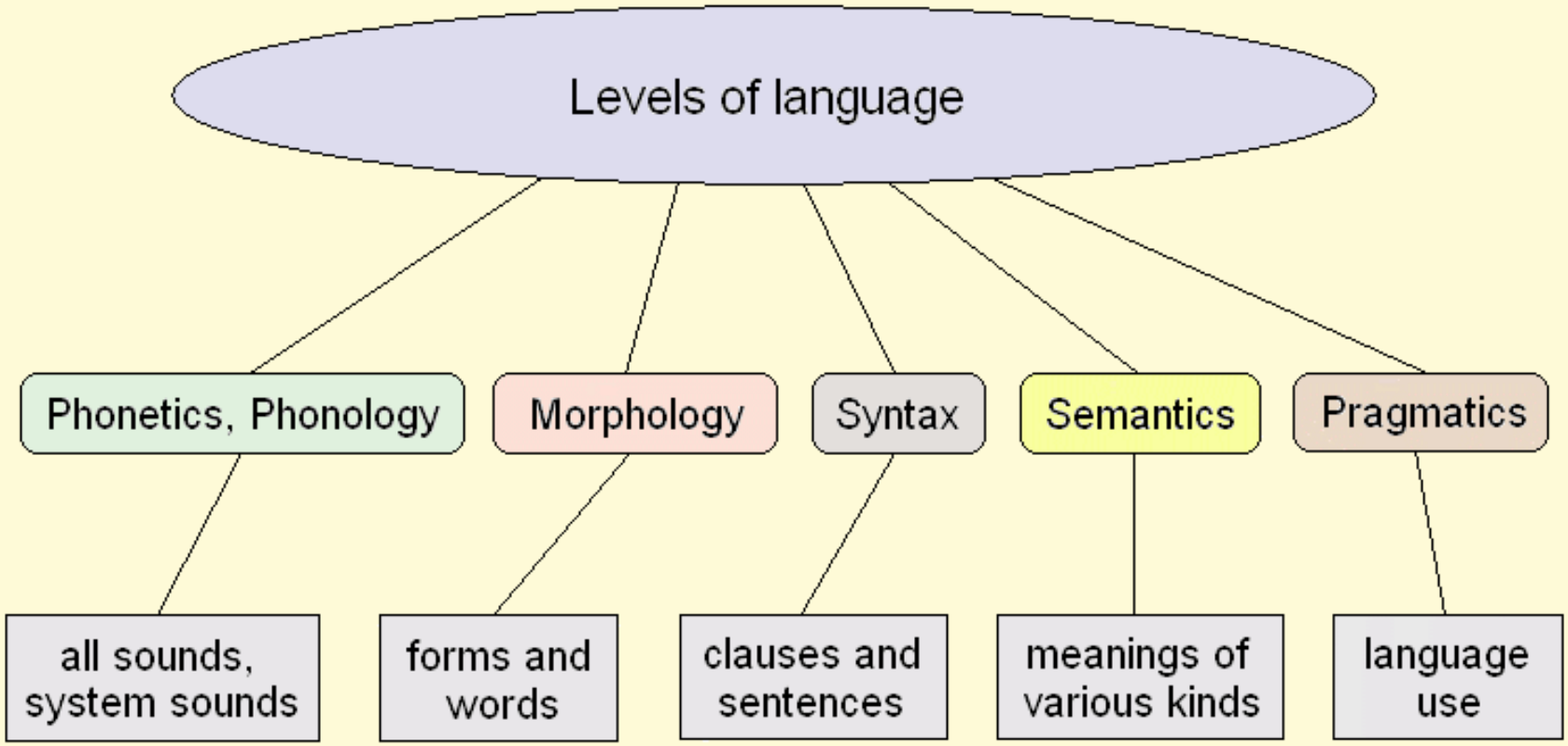
LEVELS OF LANGUAGE



LEVELS OF LANGUAGE



cdn.differencebetween.net



www.uni-due.de

Raymond Hickey
December 2005

Object of study	*Name of field*	*Size of unit*
Language use	Pragmatics	Largest
Meaning	Semantics	|
Sentences, clauses	Syntax	|
Words, forms	Morphology	|
Classified sounds	Phonology	|
All human sounds	Phonetics	Smallest ↑

Bottom-up approach to linguistic analysis

PHONETICS

(Level of Sound)

- Fonetik: himpunan bunyi manusia yang mungkin, atau menggambarkan sifat fisik dan yang membentuk suara ucapan yang terjadi dalam berbagai bahasa dunia.
- Saat kita berbicara, biasanya itu adalah satu rangkaian suara yang berkesinambungan
- Kita dapat mengidentifikasi suara ucapan ketika kita membagi serangkaian suara menjadi bit-bit yang dikenal sebagai segmen suara.
  - Misalnya, pada bahasa Inggris, kata 'cap' memiliki tiga segmen 'c', 'a', dan 'p'.
  - Ketika setiap suara diganti dengan yang lain akan menyebabkan perubahan makna. Misalnya,
    - 'cap' menjadi 'sap' ketika 'c' diganti dengan 's';
    - 'cap' menjadi 'cup' ketika 'a' diganti dengan 'u'; dan
    - 'cap' menjadi 'cab' ketika 'p' diganti dengan 'b'.
  - Kata 'sap', 'cup', dan 'cab' semuanya memiliki arti yang berbeda dalam bahasa Inggris.
- Studi fonetik menyediakan inventaris bunyi bahasa.

PHONOLOGY

(Level of Sound)

- Fonologi: mengatur dan menyusun suara untuk menyampaikan makna.
- Fonologi berkaitan dengan mengklasifikasikan bunyi bahasa dan mengatur bagaimana subset dalam bahasa tertentu digunakan,
  - misalnya perbedaan makna apa yang dapat dibuat berdasarkan bunyi apa.

MORPHOLOGY

(Level of structure)

- Morfologi struktur internal kata dan aturan pembentukan kata.
- Istilah morfologi mengacu pada analisis bentuk minimal dalam bahasa yang, bagaimanapun, terdiri dari suara dan yang digunakan untuk membangun kata-kata yang memiliki fungsi gramatikal atau leksikal.
- Morfem adalah satuan gramatikal terkecil yang bermakna. Morfem merupakan satuan dasar yang membentuk kata.

Contoh MORPHOLOGY

(Level of structure)

- Kata dasar +imbuhan
  - Imbuhan : awalan, akhiran, sisipan
  - Bahasa Indonesia: awalan (me-, ter-, di-, ber- , ….)
    - Contoh: menari, memberi, mencari, membuang, melakukan
  - Bahasa Inggris : akhiran (-ing, -s, -es, -ed, ly, …)
    - Contoh : studying, feeling, …

SYNTAX

(Level of structure)

- Sintaksis adalah struktur kalimat. Gabungan atau kombinasi beberapa kata untuk membentuk frase atau kalimat dan bagaimana kalimat tersebut ditafsirkan.
- Sintaks adalah cabang linguistik yang berkaitan dengan bagaimana kata-kata diatur untuk membangun ekspresi yang lebih panjang

Contoh SYNTAX

(Level of structure)

- Sintaks melihat aturan dan bagaimana kata-kata bisa disusun menjadi sebuah kalimat yang baik dan benar
- Contoh Syntax yang mempengaruhi makna dalam kalimat
  - Saya makan roti
  - ~~Roti makan saya~~ (salah)
- **Posisi kata dalam kalimat** Bahasa Indonesia dapat memiliki posisi Subjek, Predikat, Objek, Keterangan – kata kerja, kata sifat, kata benda
- Pemrosesan dan identifikasi posisi kata ini sering disebut dengan ***Part of Speech Tagging (POS Tagging)***.
- Bentuk kata dalam kalimat yang kita ketahui ada kata kerja (Verb), kata benda (noun), kata sifat (adjective), adverb, possessive dan lainnya.

MORFOLOGY vs SYNTAC

Morfologi	Sintaks
Ilmu yang mempelajari tentang struktur dari kata	Ilmu yang mempelajari struktur dari kalimat
Morfem/suku kata merupakan unit terkecil dari morfologi	Kata merupakan bagian terkecil dari sintaks
Mempelajari bagaimana kata dibentuk	Mempelajari susunan kata dan aturan di dalam struktur pembentukan kalimat
Morfologi terlihat sebagai kata dan berelasi dengan kata yang lain sehingga membentuk kalimat	Sintaks terlihat sebagai kalimat yang saling berelasi antara satu dengan yang lain.

SEMANTIC

(Level of Meaning)

- Semantik berkaitan dengan studi tentang makna kata, frasa, dan kalimat dalam bahasa.
- Dalam semantik, kata-kata biasanya dibagi menjadi pengertian dan referensinya. referensi suatu ekspresi adalah entitas yang dirujuknya sedangkan pengertian mengacu pada makna linguistik biasa dari suatu ekspresi.
- Contoh semantic:
  - Kakak itu menangis --- sedih/negatif

PRAGMATIC

(Level of Meaning)

- Pragmatik adalah penggunaan bahasa dalam situasi tertentu atau konteks kata/kalimat yang berhubungan erat keadaan atau situasi kata/kalimat tersebut terpaka
- Pragmatik mengungkapkan bahwa makna mempengaruhi dunia dan juga dipengaruhi oleh dunia. Ini menunjukkan bahwa makna ditentukan secara kontekstual.
- Contoh pragmatic:
  - Kakak itu menangis (di kuburan) --- sedih/negative
  - Kakak itu menangis (wisudaan) --- haru/positif
  - Ayah datang (diucapkan dengan nada datar)
  - Ayah datang! (diucapkan dengan nada tinggi)
  - Ayah datang? (diucapkan dengan nada tempo cepat)

Discourse Knowledge

- Pengetahuan tentang hubungan antar kalimat.
- Melakukan pengenalan apakah suatu kalimat yang telah dikenali mempengaruhi kalimat selanjutnya.
- Penting untuk identifikasi kata ganti orang, keterangan tempat atau aspek sementara dari informasi.
- Contoh:
  - **Ahmad** berangkat **ke sekolah**, **Ia** sedang belajar **disana**

World Knowledge

- Mencakup arti sebuah kata secara umum dan apakah arti khusus bagi suatu kata dalam suatu percakapan dengan konteks

5 DOMAINS OF LANGUAGE

By Communication Community; *source ASHA*

Phonology

The rules of speech sounds; how phonemes are used

Morphology

The rules of word structure; how morphemes are used

Syntax

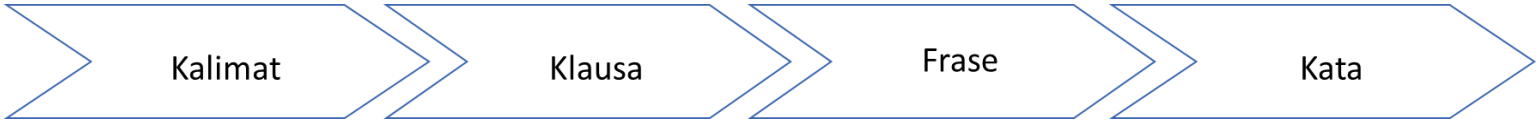
The rules of sentence structure

Semantics

The rules relating to the meaning of language

Pragmatics

The rules that occur within social situations


Kalimat
Klausa
Frase
Kata


Kalimat sesuai Tata Bahasa
Kata (Subjek)
Kata (Predikat)
Prefix
Root
Suffix
Prefix
Root
Suffix

KAMUS vs KORPUS vs THESAURUS

- **Kamus** : kata, makna
- **Korpus**: kata
  - Kamus slangword : ga = tidak
  - Korpus stopword : yang, adalah, itu, ini
- **Thesaurus** : kamus, punya banyak isian untuk kolom lain
  - Thesaurus sinonim

    ribet = ruwet, susah diatur, sulit diatur, rumit,….

BENTUK DAN STRUKTUR KATA (MORFOLOGI)

MENGOLAH KATA

- Teks dalam bahasa manusia terdiri dari kata-kata words.
- Kata memiliki berbagai informasi/attibut
  - Ejaan (orthographic spelling): bagaimana cara menuliskan sebuah kata? fox → foxes
  - Ucapan (phonetic spelling): bagaimana cara melafalkan sebuah kata? cough, bough, rough, through, . . .
  - Kelas kata (grammatical part-of-speech): bagaimana sebuah kata berinteraksi dalam konteks?
  - Makna (word sense/semantics): apakah arti sebuah kata? love : $\exists e, x, y\ love(e) \wedge lover(e, x) \wedge lovee(e, y)$

MENYIMPAN INFORMASI KATA

- Informasi kata perlu disimpan → lexicon

  Kata (word) ≈ lexical entry
- Bagaimana cara terbaik untuk menyimpannya?
  - basis data?
  - Indexing?
  - Hashtable?

Sebuah tabel besar

abdul	kt. benda	$\exists x\ abdul(x)$
ambil	kt. kerja	$\exists a, x, y\ ambil(a) \wedge ambiler(a, x) \wedge ambilee(e, y)$
⋮	⋮	⋮
zalim	kt. sifat	$\exists x\ zalim(x)$

masalah pendekatan penyimpanan

- Untuk kamus kecil, tabel mungkin masih efektif. Pada kasus
- nyata, terlalu banyak kata yang harus disimpan.
- Ada banyak hubungan dan keterkaitan antara kata.
- Nyatakan regularity - hilangkan redundancy.

Jangan disimpan sebagai entry terpisah!

- pukul
- memukul
- dipukul
- pukulan
- berpukul-pukulan

Morfologi

- Morfologi adalah ilmu yang mempelajari pembentukan kata.
- Morfem (morpheme) adalah unit bahasa terkecil yang **menyatakan makna**.
- Jenis morfem
  - **stem** / (free morpheme) / kata dasar: bisa berdiri sendiri, misal: makan, tanya, pukul
  - **affix** / (bound morpheme) / imbuhan: harus bergabung dengan stem, misal: meng-, -i, per-an
- Selain morfem, perlu seperangkat aturan yang mendefinisikan bagaimana morfem berinteraksi, misalnya: noun+-s jadi bentuk plural.

JENIS MORFOLOGI KATA

- *Agglutinative*: kata terbentuk dari gabungan banyak morfem Misalnya: bahasa Turki
- *Concatenative morphology*: kata merupakan string hasil konkatenasi morfem
- *Non-concatenative morphology*:
  - sisipan (Sunda: budak + 'ar' → barudak)
  - reduplikasi (buku + PL = buku-buku)
  - Templatic morphology: pola konsonan + variasi huruf vokal. Misalnya: Arab (morfem dasar "ktb")
    - kitåb = book
    - kutub = books
    - kåtib = writer
    - kataba = he wrote

JENIS MORFOLOGI KATA

- Beberapa metode menggabungkan morfem untuk membentuk kata:
  - Inflectional
  - Derivational
  - Compounding: merging multiple word.
    - Jerman (contoh kata pada halaman slide sebelumnya)
    - Inggris
      - web + site → website
      - home + work → homework
- Cliticization
  - Indonesia (kupinjam bukunya)
  - Inggris (he'd do his homework)

Inflectional Morphology

- Ciri-ciri Inflectional morphology
  - Sistematis: polanya teratur, maksud dan hasilnya jelas
  - Produktif: dapat diterapkan pada semua kata dengan category yang sesuai
  - Kata baru memiliki category yang sama
- Contoh: membuat plural noun dengan +s.
  - dog+s = dogs
  - fox+s = foxes
  - blog+s = blogs

Derivational Morphology

- Ciri-ciri Derivational morphology
  - Tidak sistematis: maksud dan hasil bisa berbeda
  - Tidak produktif: belum tentu bisa diterapkan pada semua kata
  - Kata baru cenderung memiliki category yang beda
- Contoh: membuat kata kerja dengan + ise / + ize.
  - national+ize = nationalize
  - incentive+ize = incentivize
  - book+ize = bookize ??

Beberapa Isu Morfologi

- Pengecualian aturan
  - looking = look + ing
  - rethink = re + think
  - thing = th + ing ??
  - read = re + ad ??
- Kerancuan: ada beberapa cara merangkai morfem
  - "undoable": ["un"+ "do"] + "able", "un"+ ["do" + "able"]
  - "penanya": "pena"+ "-nya", "pe-" + "tanya"
- Variasi ejaan
  - pray + s = prays
  - candy + s = candies
  - goose + s = geese

Morfologi dalam Berbagai Bahasa

Turki

- uygarlaştıramadıklarımızdanmışsınızcasına

uygar	civilized
laş	+Bec
tır	+Caus
ama	+NegAble
dık	+PPart
lar	+PL

ımız	+P1PL
dan	+Abl
mış	+Past
sınız	+2PL
casına	+Aslf

(behaving) as if you are among those whom we could not civilize/cause to become civilized

Jerman

- Rechtsschutzversicherungsgesellschaften
  insurance companies providing legal protection

Morfologi dalam Berbagai Bahasa

Rusia

- zhenshina devochke dala knigu
  woman+NOM girl+DAT gave book+ACC
  'the woman gave the girl a book'

- zhenshine devochka dala knigu
  woman+DAT girl+NOM gave book+ACC
  'the girl gave the woman a book'

Indonesia

- mengotak-ngotakkan
- lauk-pauk