Carnegie Mellon University
Language
Technologies
Institute

11-411/11-611 Natural Language Processing

Words, Morphologies, and Lexicons

David R. Mortensen

January 26, 2023

Language Technologies Institute

Words Have Meaningful Parts

If You Want to Start a Brawl in a Bar Full of Linguists...

...shout, “I have discovered a universal definition for ‘word’!”

It Depends on How You Divide Them

Text can be divided into sequences of words (tokenized, in NLP terms) according to various criteria:

- **Orthographically** (e.g., splitting on white space and punctuation)
- **Phonologically** (e.g., splitting text into tokens that satisfy certain conditions on pronunciation)
- **Syntactically** (e.g., splitting text into tokens that serve as units in the grammar)

And others.

Even for the same language, these do not always produce the same results. For example, the *'s* in *student's* is orthographically part of the preceding word and acts, from the standpoint of sound, like other suffixes (e.g., the -*s* in *books*. However, syntactically it is a separate unit.

Words Have Meaningful Parts

However you define words, some words cannot be disassembled. Take giraffe. Sure, you can divide it up letter by letter

g-i-r-a-f-f-e

or sound-by-sound

d͡ʒ-ə-ɹ-æ-f

But these units are meaningless, by themselves. They are GRAPHEMES and PHONEMES. *Giraffe* only has one meaningful part.

Even English and Chinese Words Can Have Internal Structure

English and Chinese (among other languages) have lots of words with only one meaningful unit—one morpheme. But other words have many morphemes (meaningful parts):

reoperationalizations → re-operat(e)-(t)ion-al-iz(e)-ation-s

These words have **internal structure**

Chinese Words, too, Can Have Internal Structure

Even in Chinese, which is famous for one-morpheme words, words can have internal structure:

我	‘I’	我们	‘we’
你	‘you’	你们	‘ya’ll’
他	‘he’	他们	‘they’
同志	‘comrade’	同志们	‘comrades’

Morphology is the Study of the Internal Structure of Words

Morphology is the study of the internal structure of words

- How morphemes combine
- How morphemes function

Dividing Words into Morphemes is Morphological Segmentation

Dividing words into morphemes is called morphological segmentation. How do you do it?

Look for substrings that correspond to units of meaning. In the Mandarin Chinese example, all the words in the second column have a plural meaning, and all of them end in 们, so we can conclude that 们 is probably a morpheme meaning “plural”.

This is confirmed, because when we take 们 away, what is left over is also meaningful. For example, 同志 means ‘comrade’ by itself.

Most Languages have Morphologies More Complicated than English and Chinese

English and Chinese have very simple morphologies. Most languages have much more complicated word structures. An unsolved challenge of NLP is dealing effectively with morphologically complex languages.

NLP Practicioners Should Care about Morphology

- Morphology makes instances of the same word look like different words, leading to data sparcity
- In morphologically rich languages, single words of many morphemes may express concepts and relationships that are expressed by a full phrase or sentence in languages like English or Chinese, calling for CHARACTER-LEVEL MODELS, SUBWORD TOKENIZATION, MORPHOLOGICAL SEGMENTATION, or MORPHOLOGICAL ANALYSIS
- Generating text in morphologically complex languages is complicated, particularly when translating from morphologically impoverished languages that distribute information very differently

Morphology and Subword Tokenization (Morfessor, BPE, and FSTs)

Turkish: “And he killed James the brother of John with the sword”

Segmentation	Example
Tokenized	Yuhannanın kardeşi Yakubu kılıçla öldürdü .
Character	Y u h a n n a n ı n _ k a r d e ş i _ Y a k u b u _ k ı l ı ç l a _ ö l d ü r d ü .
BPE (see next)	**Yuhan@@ nanın kardeşi Yakubu kılıçla öldürdü .**
Morfessor	Yuhanna@@ nın kardeş@@ i Yakub@@ u kılıç@@ la öldürdü .
FST+BPE	Yuhan@@ nanın kardeş@@ i Yakub@@ u kılıç@@ la öl@@ dür@@ dü .
FST+Morfessor	**Yuhanna@@ nın kardeş@@ i Yakub@@ u kılıç@@ la öl@@ dür@@ dü .**

Adapted from Park et al. (2021).

BPE is Byte-Pair Encoding

- Way of dividing words into subword tokens
- Divides words to minimize one-off tokens (hapax legomena)

Start	Step 1	Step 2	Step 3
aaabdaaabac	ZabdZabac	ZYdZYac	XdXac
	Z=aa	Y=ab	X=ZY
		Z=aa	Y=ab
			Z=aa

And so on, recursively until a vocabulary size is reached.

BPE sometimes yields morpheme-like segments, but often does not

BPE Operates Blindly

	peed	*deed*
Linguist_1	pe@@ ed	deed
Linguist_2	pee@@ d	deed
BPE_1	pe@@ ed	de@@ ed
BPE_2	peed	deed

Roots and Affixes

Roots and Affixes are Kinds of Morphemes

Some morphemes give words their basic meaning. These are called ROOTS. Other morphemes are added to words to make new words, or to make new forms of existing words.

un-think-able; kitten-s

These are called AFFIXES.

Roots Provide Basic Meaning and Affixes Modify It

- The roots (in the first column) express the basic meaning
- Affixes add grammatical meaning (2nd column) or modify the semantic meaning (3rd column)

<root>	<root>ing	<root>er
run	running	runner
think	thinking	thinker
program	programming	programmer
kill	killing	killer

Bound Morphemes Do not Occur as Independent Words

Some roots are BOUND, that is, they don’t occur as independent words. Consider –ceive in

- conceive
- deceive
- receive
- perceive

These relationships are sometime opaque: etymology

Four Types of Affixes Attach to Bases

- **Bases** are strings to which affixes can be applied
- **Prefixes** are added to the beginning of a base
- **Suffixes** are added to the end of the base
- **Circumfixes** are added the beginning and end of the base simultaneously
- **Infixes** are inserted inside the base

Prefixes are Added Before the Base; Suffixes are Added After It

Prefixes	regular		→	**ir**-regular	
	nuptial	‘wedding’	→	**pre**-nuptial	‘before wedding’
Suffixes	real	‘actual’	→	real-**ize**	‘become actual’
	hunt		→	hunt-**er**	

Circumfixes Are Added Around the Base while Infixes Are Inserted Inside of It

Circumfix	sammel[*]	‘gather’	→	**ge**-sammel-**t**	‘has gathered’
	light	‘light’	→	**en**-light-**en**	‘make light’
	bold	‘bold’	→	**em**-bold-**en**	‘make bold’
Infix	California		→	Cali-**freakin’**-fornia	
	sulat[†]	‘write!’	→	s**um**ulat	‘will write’

[*] From German, a major language of Europe

[†] From Tagalog, a major language of the Philippines

There Are Non-Concatenative Morphological Operations

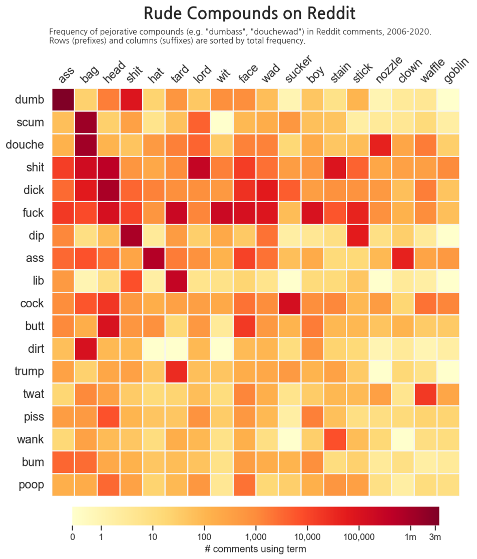
There are morphological operations on bases other than affixation:

- Reduplication
  - sulat* → susulat
  - anak[†] ‘child’ → anakanak ‘children’
- Apophony (umlaut, ablaut, etc)
  - foot → feet
  - sing → sang, sung
- Transfixation (root-and-pattern or templatic morphology)
  - ktb → kitab[‡] ‘book’, kataba ‘he wrote’, taktub ‘she writes’, etc.

*Tagalog. [†]Indonesian. [‡]Arabic.

Compounding

English Speakers Make New Compounds Constantly



Rude Compounds on Reddit

Frequency of pejorative compounds (e.g. "dumbass", "douchewad") in Reddit comments, 2006-2020.
Rows (prefixes) and columns (suffixes) are sorted by total frequency.

Inflection and Derivation

Derivational Morphology Creates New Words while Inflectional Morphology Adds Information to Existing Words

- Some affixes (and other morphological operations) create new words by changing meaning or part of speech: **derivation**
- Some operations merely add information to a word based on the grammatical context in which it occurs: **inflection**

Examples of English Derivation

- The *-er* suffix converts verbs into nouns that do the action encoded by the verb: *hunter*, *shooter*, *killer*, *fighter*, *peacemaker*
- It changes both the meaning and part of speech of a base; therefore, it must be derivational
- The *un-* prefix negates the meaning of an adjective (*unfriendly*, *undefined*, *unfathomable*), but does not change its part of speech
- It is also derivational.

Examples of English Inflection

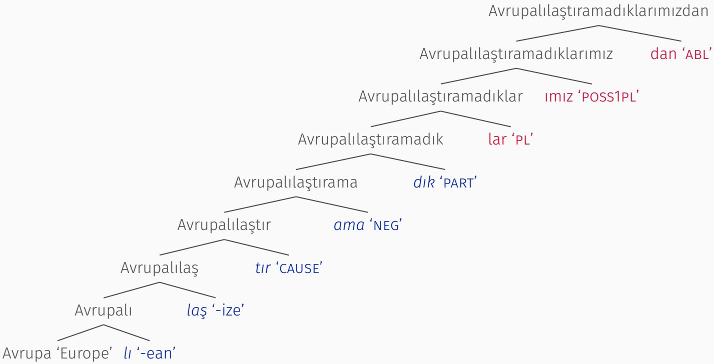
- The *-s*/*-es* suffix indicates that nouns are plural: *consonants*, *vowels*, *morphemes*, *lambdas*
- It does not create new words (change meaning or part-of-speech); it just MARKS nouns as having a particular property—plurality
- The *-ing* suffix indicates that verbs are "progressive": *coding*, *running*, *training*, etc.
- It is inflectional (in these cases)
- However, when *-ing* makes verbs into nouns (*a build**ing***) then it is derivational.

Derivational Affixes Tend to Occur Closer to the Root than Inflectional Affixes

- Words can have both inflectional and derivational affixes at the same time
- Derivational affixes tend to appear closer to the root
- Inflectional suffixes tend to appear farther from the root (on the outer margins of the word)
- This turns out to be convenient

Turkish Provides a Good Example of Affix-Ordering

Turkish: ‘of ours that were unable to be Europeanized.’ Blue morphemes are derivational; Red morphemes are inflectional



Avrupalılaştıramadıklarımızdan

Avrupalılaştıramadıklarımız dan ‘ABL’

Avrupalılaştıramadıklar ımız ‘POSS1PL’

Avrupalılaştıramadık lar ‘PL’

Avrupalılaştırama *dık* ‘PART’

Avrupalılaştır *ama* ‘NEG’

Avrupalılaş *tır* ‘CAUSE’

Avrupalı *laş* ‘-ize’

Avrupa ‘Europe’ *lı* ‘-ean’

It Is Often Desirable to Get Rid of Inflectional Morphemes while Retaining Derivational Morphemes

- **Lemmatization** — return lemma (”dictionary” form of word)
- **Stemming** — “chop off” morphemes until a stem-like string remains

In this course, we will explore one approach to lemmatization (Homework 2). Using a lemmatizer may be useful for you project. For your project, you may also want to employ stemming.

Modeling How Morphemes Combine

Morphemes Are Selective about What They Attach to

- The study of how morphemes combine in sequence is called **morphotactics**
- Why can you say *thankfulness* but not *thankfulity*?
- Why can you say *prodigious* and *vitalize* but not *prodigiousize*?

Morphotactics Is about More than Just Memorizing Combinations of Morphemes

- **Productivity**: morphology allows speakers and writers to make new words out of old (and new) parts
- If someone told you they were going to *gorbalize* your homework, you would also suspect that *gorbalization* of homework was possible (because you can form nouns from verbs ending in -*ize* by adding -*ation*)

There Are Selectional Restrictions in Morphology

- **Linguistics** — there are principled limits on what kinds of base an affix will attach to
- **NLP** — what can come after is contingent on what comes before
- *nation-al-ize*, *nation-al-iz-ation*, and *con-feder-ate* but not **con-feder-ate-ation*

Modeling Allomorphic Alternation

A Morpheme Can Have More than One Shape (Spelling or Pronunciation) Depending on the Environment in which It Occurs

Isolation	Before -er	Before -ing
fib	fibb-er	fibb-ing
bid	bidd-er	bidd-ing
squig	squigg-er	squigg-ing

The Two -*s* Suffixes in English Take the Same Forms

- Two “-s” suffixes:
    - Plural (nouns)
    - 3rd person singular non-past (verbs)
- Both have -s and -*es* forms

-s		-es	
pick	picks	watch	watches
laugh	laughs	fish	fishes
waif	waifs	pass	passes
pin	pins	kiss	kisses
pill	pills	fizz	fizzes
pew	pews	pox	poxes
bay	bays	sax	saxes

We Can Formulate a Generalization and a Rule

To a first approximation:

- -es occurs after s, x, z, ch, and sh
- -s occurs elsewhere

Rule:

- Start with -^s (^ is a morpheme boundary)
- Insert -e- between ^ and s when ^ is preceded by s, x, z, ch, or sh

Getting the Right Allomorphs by Applying Two Rules in Sequence

From Lexical Form to Surface Form (Output)

Lexical Form	pin +Singular	pin +Plural	pass +Singular	pass +Plural
Input / Rules	pin	pin^s	pass	pass^s
Rule 1	pin	pin^s	pass	pass^es
Rule 2	pin	pins	pass	passes
Surface Form	**pin**	**pins**	**pass**	**passes**

Questions?