CS11-737 Multilingual NLP

Words and Morphology

Lei Li

https://lileicc.github.io/course/11737mnlp23fa/

Carnegie Mellon University
Language Technologies Institute

originally by Yulia Tsvetkov and Alan Black

What is a word?

- How many words?

  Bob’s handyman is a do-it-yourself kinda guy, isn’t he?

What is a word?

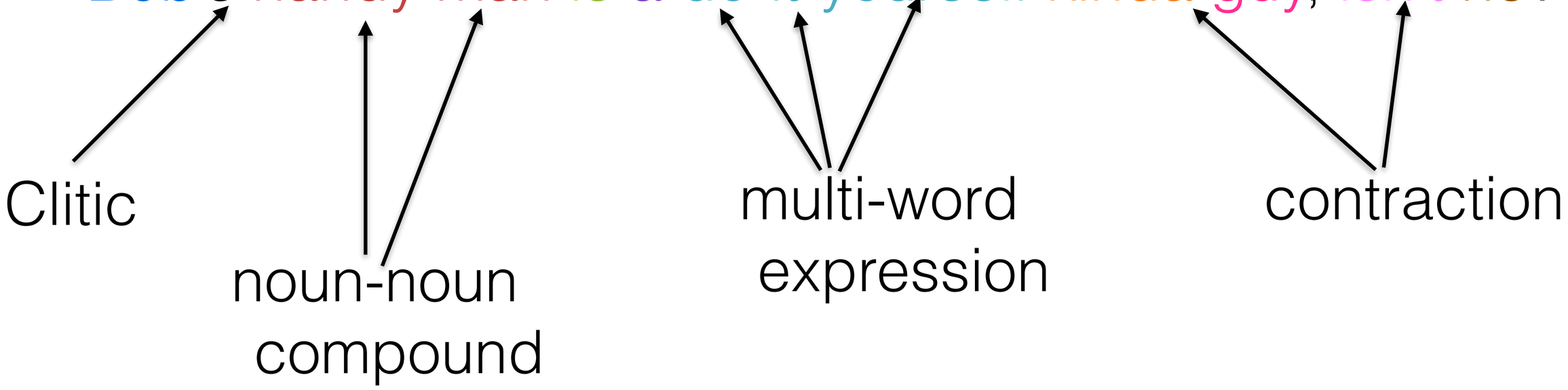
- How many words?

  Bob’s handyman is a do-it-yourself kinda guy, isn’t he?

What is a word?

- How many words?

Bob's handy man is a do-it-yourself kinda guy, isn't he?



What is a word?

- How many words?

Bob's handyman is a do-it-yourself kinda guy, isn't he?

Much'ananayakapushasqakupuniñataqsunamá

Much'a -na -naya -ka -pu -sha -sqa -ku -puni -ña -taq -suna -má

"So they really always have been kissing each other then"

```
Much'a   to kiss
-na      expresses obligation, lost in translation
-naya    expresses desire
-ka      diminutive
-pu      reflexive (kiss *eachother*)
-sha     progressive (kiss*ing*)
-sqa     declaring something the speaker has not personally witnessed
-ku      3rd person plural (they kiss)
-puni    definitive (really*)
-ña      always
-taq     statement of contrast (...then)
-suna    expressing uncertainty (So...)
-má      expressing that the speaker is surprised
```

(example from Quechua)

Turkish	English
kork(-mak)	(to) fear
korku	fear
korkusuz	fearless
korkusuzlaş (-mak)	(to) become fearless
korkusuzlaşmış	One who has become fearless
korkusuzlaştır(-mak)	(to) make one fearless
korkusuzlaştırıl(-mak)	(to) be made fearless
korkusuzlaştırılmış	One who has been made fearless
korkusuzlaştırılabil(-mek)	(to) be able to be made fearless
korkusuzlaştırılabilecek	One who will be able to be made fearless
korkusuzlaştırabileceklerimiz	Ones who we can make fearless
korkusuzlaştırabileceklerimizden	From the ones who we can make fearless
korkusuzlaştırabileceklerimizdenmiş	I gather that one is one of those we can make fearless
korkusuzlaştırabileceklerimizdenmişçesine	As if that one is one of those we can make fearless
korkusuzlaştırabileceklerimizdenmişçesineyken	when it seems like that one is one of those we can make fearless

Structural Subfields of Linguistics

Phonetics	The study of the sounds of human language
Phonology	The study of sound systems in human languages
Morphology	The study of the formation and internal structure of words
Syntax	The study of the formation and internal structure of sentences
Semantics	The study of the meaning of sentences
Pragmatics	The study of the way sentences with their semantic meanings are used for particular communicative goals

Words

- Orthographic definition
  - strings separated by white spaces
  - spoken language: units corresponding to written word separated by pause
  - problem: Bob's handy man is a do-it-yourself kinda guy, isn't he?
- What about languages that do not use white spaces?

他昨天晚上去看了消失的她

he yesterday night watched lost in stars

- Unwritten languages

Words

- Prosodic definition
  - words have one main stress and longer words may have a secondary stress
  - problems: function words, clitics

Words

- Syntactic definition:
  - words are the syntactic building blocks of sentences
- Semantic definition
  - words are units that describe a single idea or a semantic concept
  - problem: many semantic concepts span phrases or sentences and don't have a corresponding word

Parts of Speech

- Open classes
  - nouns
  - verbs
  - adjective
  - adverbs

Adj (JJ)
Adv
Conjunction
DT
Noun
Number
Prep (IN)
Pronoun
Verb

- Closed classes
  - prepositions
  - determiners
  - pronouns
  - conjunctions
  - auxiliary verbs

Annette	has	written	two	artificial	intelligence	policies	for	her	University	of	Pittsburgh	colleagues	to	consider	including	in	their	syllabi	.
NN	V	V	CD	JJ	NN	NN	IN	PPZ	NN	IN	NN	NN		V	V	IN	PPZ	NN	.

Part of speech tags

POS Tag	Description	Example
CC	coordinating conjunction	and
CD	cardinal number	1, third
DT	determiner	the
EX	existential there	there is
FW	foreign word	les
IN	preposition, subordinating conjunction	in, of, like
IN/that	that as subordinator	that
JJ	adjective	green
JJR	adjective, comparative	greener
JJS	adjective, superlative	greenest
LS	list marker	1)
MD	modal	could, will
NN	noun, singular or mass	table
NNS	noun plural	tables
NP	proper noun, singular	John
NPS	proper noun, plural	Vikings
PDT	predeterminer	both the boys
POS	possessive ending	friend's
PP	personal pronoun	I, he, it
PPZ	possessive pronoun	my, his
RB	adverb	however, usually, naturally,
RBR	adverb, comparative	better
RBS	adverb, superlative	best
RP	particle	give up
SENT	Sentence-break punctuation	. ! ?
SYM	Symbol	/ [= *
TO	infinitive 'to'	togo
UH	interjection	uhhuhhuhh
VB	verb be, base form	be
VBD	verb be, past tense	was, were
VBG	verb be, gerund/present participle	being
VBN	verb be, past participle	been
VBP	verb be, sing. present, non-3d	am, are
VBZ	verb be, 3rd person sing. present	is
VH	verb have, base form	have
VHD	verb have, past tense	had
VHG	verb have, gerund/present participle	having
VHN	verb have, past participle	had
VHP	verb have, sing. present, non-3d	have
VHZ	verb have, 3rd person sing. present	has
VV	verb, base form	take
VVD	verb, past tense	took
VVG	verb, gerund/present participle	taking
VVN	verb, past participle	taken
VVP	verb, sing. present, non-3d	take
VVZ	verb, 3rd person sing. present	takes
WDT	wh-determiner	which
WP	wh-pronoun	who, what
WP$	possessive wh-pronoun	whose
WRB	wh-abverb	where, when
#	#	#
$	$	$
"	Quotation marks	"
``	Opening quotation marks	"
(	Opening brackets	({
)	Closing brackets	) }
,	Comma	,
:	Punctuation	– ; : — …

Chinese Part-of-Speech Tags

Tag	Description	Example
AD	adverb	也
AS	aspect marker	着
BA	把 in ba-construction	把
CC	coordinating conjunction	和
CD	cardinal number	一百
CS	subordinating conjunction	虽然
DEC	的 in a relative-clause	的
DEG	associative	的
DER	in V-de const. and V-de-R	得
DEV	地 before VP	地
DT	determiner	这
ETC	for words 等, 等等	等, 等等
FW	foreign words	A
IJ	interjection	哈哈
JJ	other noun-modifer	新
LB	被 in long bei-const	被
LC	localizer	里
M	measure word	个

Tag	Description	Example
MSP	other particle	所
NN	common noun	工作
NR	proper noun	中国
NT	temporal noun	目前
OD	ordinal number	第一
ON	onomatopoeia	
P	Prepositions (excluding 把 and 被)	在
PN	pronoun	我
PU	punctuation	标点
SB	被 in short bei-const	被
SP	sentence-final particle	吗
VA	predicative adjective	好
VC	copula	是
VE	有 as the main verb	有
VV	other verbs	要
X	numbers and units, mathematical sign	59mm

The Universal Dependencies

Universal Dependencies (UD) is a framework for consistent annotation of grammar (parts of speech, morphological features, and syntactic dependencies) across different human languages. UD is an open community effort with over 300 contributors producing nearly 200 treebanks in over 100 languages. If you're new to UD, you should start by reading the first part of the Short Introduction and then browsing the annotation guidelines.

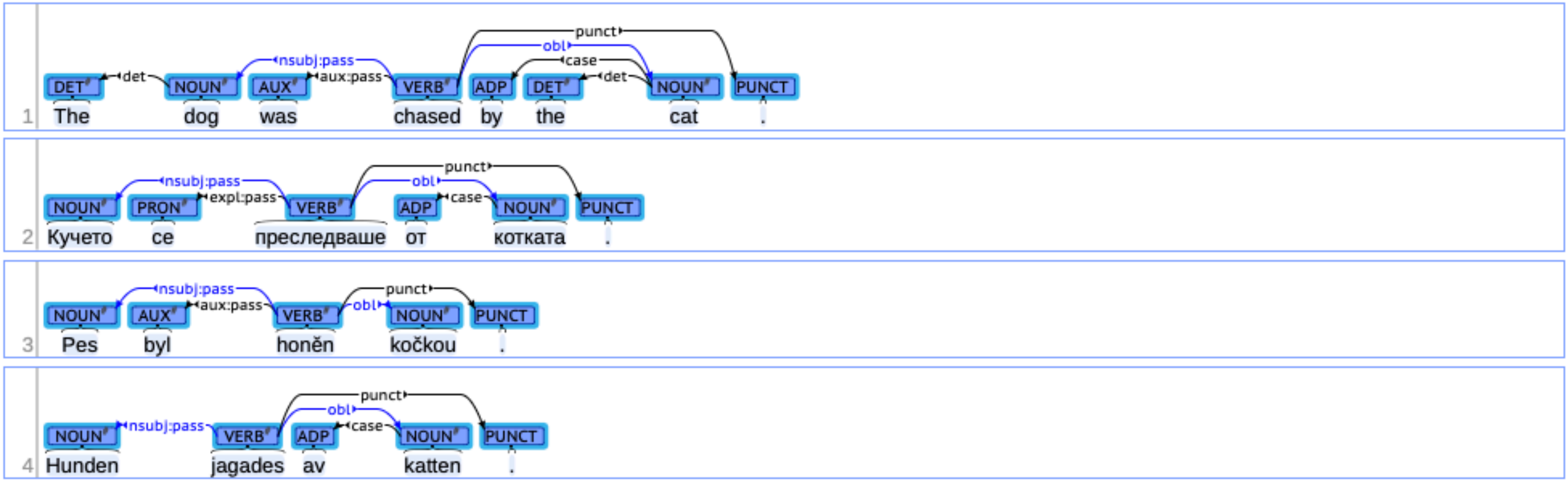
- Short introduction to UD
- UD annotation guidelines
- More information on UD:
  - How to contribute to UD
  - Tools for working with UD
  - Changes to the UD guidelines
  - UD-related events
  - Projects related to UD
- Query UD treebanks online:
  - PML Tree Query maintained by the Charles University in Prague
  - TEITOK maintained by the Charles University in Prague
  - Grew-match maintained by Inria in Nancy
  - INESS maintained by the University of Bergen
- Download UD treebanks

Open class words	Closed class words	Other
ADJ	ADP	PUNCT
ADV	AUX	SYM
INTJ	CCONJ	X
NOUN	DET	
PROPN	NUM	
VERB	PART	
	PRON	
	SCONJ	

https://universaldependencies.org/

The Universal Dependencies

- Example



Morpheme

- A meaningful morphological unit of a language that can not be further divided
- e.g.
  - disregard
  - kindness

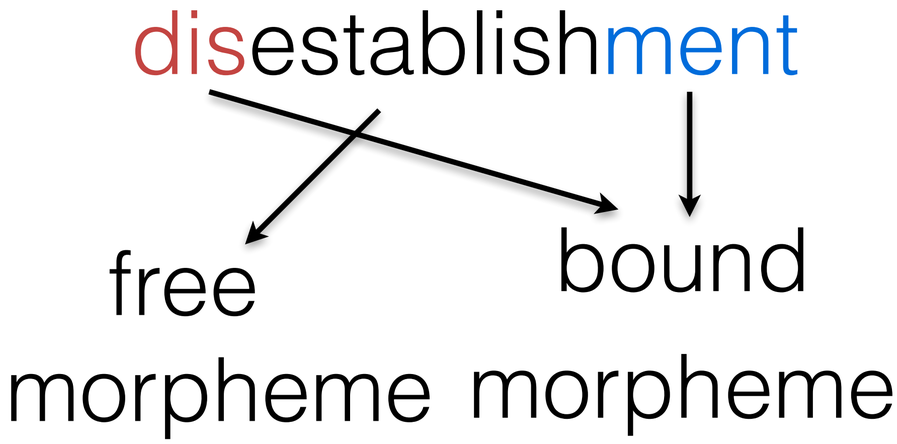
establish (V)

disestablish (V)

disestablishment (N)

antidisestablishment (N)

antidisestablishmentary (A)



Morphological processes

- Concatenation
- affixation=stem+affix
  - prefix
  - suffix
- non-concatenative affixation
  - infix
- compounding = stem + stem

establish (V) —> stem

disestablish (V) —> prefix + stem

disestablishment (N) —> prefix + stem + suffix

antidisestablishment (N)

antidisestablishmentary (A)

dish (N)+washer (N) = dishwasher (N)

Morpheme in Chinese

- Simple word:
  - 人 (human)
  - 葡萄 (grape)
  - 蝴蝶 (butterfly)
  - 沙发 (sofa, loan word)
  - 轰隆隆 (sound of thunder, onomatopoetic word)
- compound word
  - 老师 (old teacher = teacher)
  - 现代化 (modernization, —化)
  - 日出（sun rise, subject-predicate）
  - 打篮球（play basketball, verb-object
  - 黑板 （blackboard）
  - 证明（prove）
  - 矛盾 （controversy）
  - 洗衣机（wash cloth machine）
  - 妈妈 （mom）

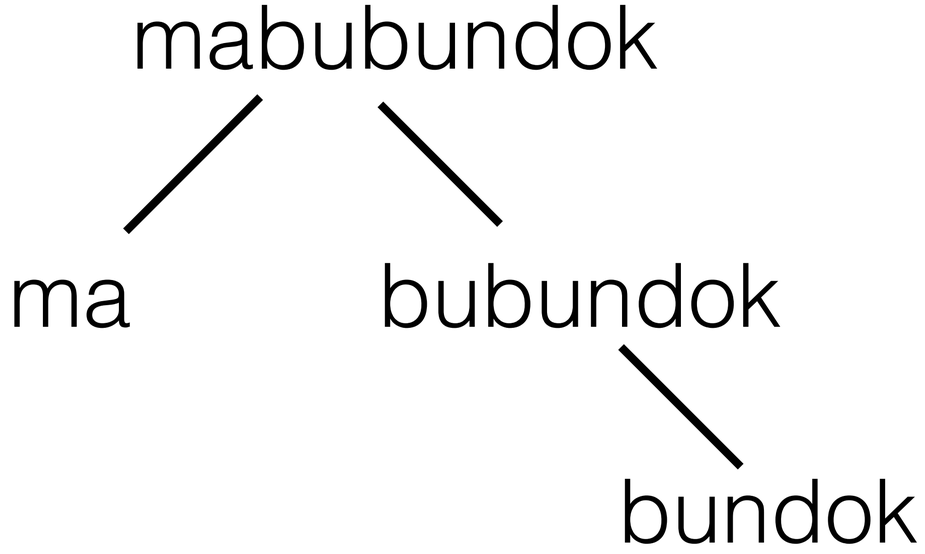
Arabic

- Root and pattern morphology

katab-a	he wrote
kaataba	he corresponded
kutib-a	it was written
kitaab	book
kutub	books
kaatib	writer; writing
kuttaab	writers
uktub	write (to a male)

Tagalog

- stem - bundok
- singular - mabundok
- plural - mabubundok
- gloss - “mountainous”



Morphological functions

- Derivational morphemes
  - bound morphemes used to create new words
  - if these affixes are attached to a new base, the resulting combination yields a word with a new meaning
  - often derived word belongs to a different syntactic class

establish (V)

disestablish (V)

disestablishment (N)

- Inflectional morphemes
  - bound morphemes used to mark grammatical distinctions
  - change the form but not POS tag or the key meaning of the word

grow

grows

Morphological Levels

- Morphosyntax
  - how stems and affixes combine
  - e.g. verb + ed, verb + ing, un-grace-ful-ly
- Morphophonemics
  - pronunciations/orthographic modifications at boundaries
  - “e” gets deleted when preceded by a consonant, and followed by a morpheme boundary and morpheme starting with e
  - e.g. cooked
  - “n” becomes “m” at morpheme boundary followed by “m”, “b”, “p”
  - morphophonemics can make morphology non-segmental

Morphological typology

- Isolating or analytic
  - Vietnamese, Chinese, English
- Synthetic
  - Fusional or Flexional
    - German, Greek, Russian
    - Templatic: Hebrew and Arabic
  - Agglutinative or Agglutinating
    - Finnish, Turkish, Malayalam, Swahili
  - Polysynthetic
    - Inuit, Yupik

UniMorph

- The Universal Morphology (UniMorph) project is a collaborative effort to improve how NLP handles complex morphology in the world's languages. The goal of UniMorph is to annotate morphological data in a universal schema that allows an inflected word from any language to be defined by its lexical meaning, typically carried by the lemma, and by a rendering of its inflectional form in terms of a bundle of morphological features from our schema. The specification of the schema is described here and in Sylak-Glassman (2016).
- 169 languages

```
pip install unimorph
```

https://unimorph.github.io/

SIGMORPHON

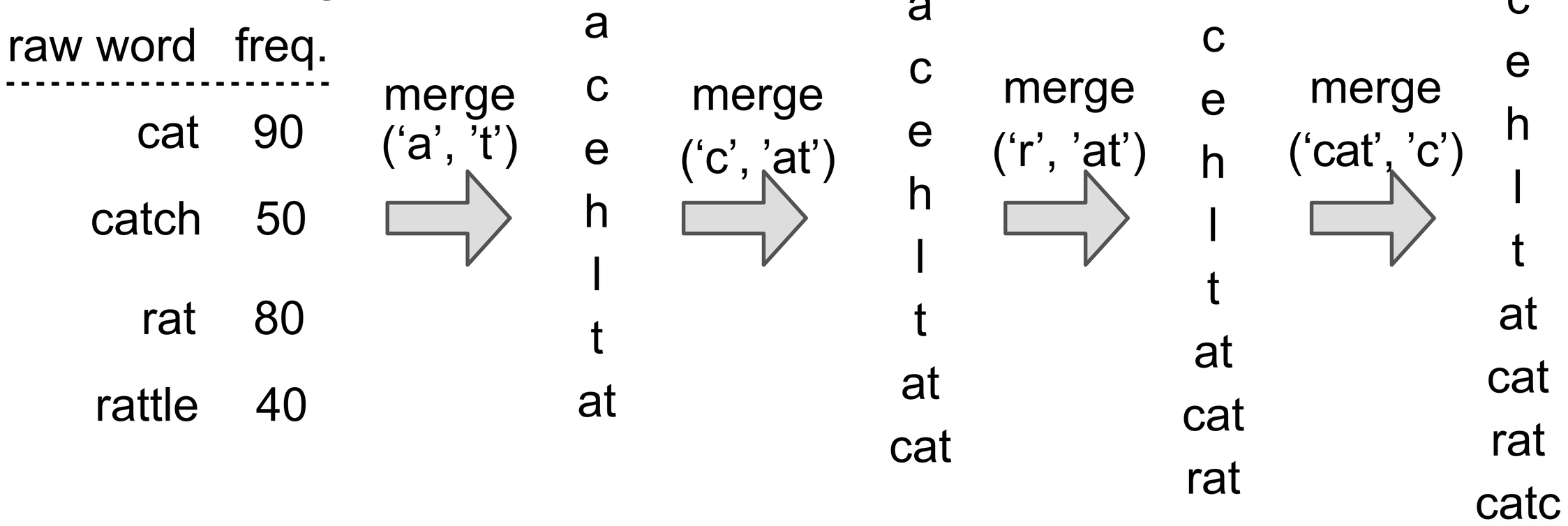
- Usually co-located with ACL
- Shared tasks
  - Cross-lingual transfer for morphological inflection
  - Morphological analysis in context
  - Morphological paradigm completion

Morphological Analyzers

- Finite state morphology
  - skilled, but not very hard (by experts)
  - Xfst, FOMA
- Unsupervised methods
  - Morfessor
  - Assumes segmental view of morphology
- Stemming
  - remove end of words
- Byte-pair-encoding (BPE)
  - not necessary semantic meaningful, but statistical segmental splits

Byte-Pair-Encoding Tokenization

- Byte-Pair-Encoding (BPE)
  - starting from chars
  - repeatedly, merge most frequent pairs to form new tokens
  - until reaching a fixed size.



Neural Machine Translation of Rare Words with Subword Units. Sennrich et al. ACL 2016

Related NLP Problems

- Tokenization
- Lemmatization
- Text normalization
  - replace numbers, symbols, abbreviations with standard words
- Spelling correction/grammatical error correction
- Processing words in multilingual NLP tasks, e.g. language modeling or machine translation
- syntactic tagging (next class) and morphological analysis
- Evaluation of text generation or machine translation (of on the word level)