From Words to Linguistic Structure

Sandeep Soni

04/06/2024

Words

They can fish

Words

They can fish

Is this sentence about the ability to catch fish?

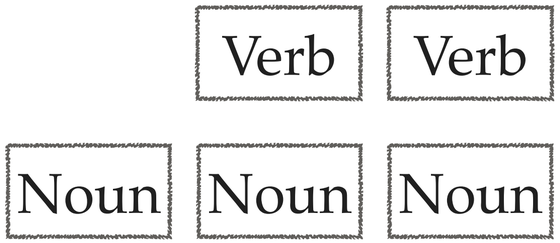
Words

They can fish

… Or is the sentence about storing fish?

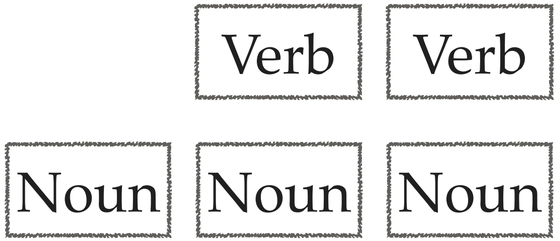
Parts of Speech

Categories of words that have similar morphological and syntactic properties



Contextual dependence

To find the syntactic role that a word has in a sentence, more context is helpful



They can fish

They can fish in a cold room

Part of Speech tagging

The task of assigning a tag (category) to every word in a sequence.

TAG SET

The number of categories depends on language, genre, etc although there are some core categories

They can fish in a cold room

Nouns	People, places, things, dates, etc. Depend on quantity (singular Vs Plural)
Adjectives	Properties or qualities of noun phrases
Verbs	Actions, processes. Depends on tense, number, person, etc
Adverbs	Modifiers of verbs; qualify the action (e.g., quickly stopped)
Determiner	Beginning of a noun phrase; emphasizes specificity (e.g., a film)
Prepositions	Indicates spatial/direction/temporal relationship with a noun phrase (e.g., in a room)
Conjunctions	Connectives between phrases, clauses, sentences (e.g., Jack and Jill)
Pronouns	References to noun phrases (e.g., he, she, it)

Tag	Description	Examples	%
Nominal, Nominal + Verbal			
N	common noun (NN, NNS)	books someone	13.7
O	pronoun (personal/WH; not possessive; PRP, WP)	it you u meeee	6.8
S	nominal + possessive	books' someone's	0.1
^	proper noun (NNP, NNPS)	lebron usa iPad	6.4
Z	proper noun + possessive	America's	0.2
L	nominal + verbal	he's book'll iono (= *I don't know*)	1.6
M	proper noun + verbal	Mark'll	0.0
Other open-class words			
V	verb incl. copula, auxiliaries (V*, MD)	might gonna ought couldn't is eats	15.1
A	adjective (J*)	good fav lil	5.1
R	adverb (R*, WRB)	2 (i.e., *too*)	4.6
!	interjection (UH)	lol haha FTW yea right	2.6

Other closed-class words			
D	determiner (WDT, DT, WP$, PRP$)	the teh its it's	6.5
P	pre- or postposition, or subordinating conjunction (IN, TO)	while to for 2 (i.e., *to*) 4 (i.e., *for*)	8.7
&	coordinating conjunction (CC)	and n & + BUT	1.7
T	verb particle (RP)	out off Up UP	0.6
X	existential *there*, predeterminers (EX, PDT)	both	0.1
Y	X + verbal	there's all's	0.0
Twitter/online-specific			
#	hashtag (indicates topic/category for tweet)	#acl	1.0
@	at-mention (indicates another user as a recipient of a tweet)	@BarackObama	4.9
~	discourse marker, indications of continuation of a message across multiple tweets	RT and : in retweet construction RT @user : hello	3.4
U	URL or email address	http://bit.ly/xyz	1.6
E	emoticon	:-) :b (: <3 o_O	1.0

Miscellaneous			
$	numeral (CD)	2010 four 9:30	1.5
,	punctuation (#, $, '', (,), ,, ., :, ``)	!!! .... ?!?	11.6
G	other abbreviations, foreign words, possessive endings, symbols, garbage (FW, POS, SYM, LS)	ily (*I love you*) wby (*what about you*) 's ♫ --> awesome...I'm	1.1

Source: Gimpel et. al.

Sequence Labeling

- Classification: Predict the tag for each word independently

The old man the boat

Sequence Labeling

- Structured prediction: Predict the entire tag sequence

$$f([x_1, x_2, \ldots, x_n]) \rightarrow [y_1, y_2, \ldots, y_n]$$

Where can part of speech tagging be used?

Sarcasm as Contrast between a Positive Sentiment and Negative Situation

Ellen Riloff, Ashequl Qadir, Prafulla Surve, Lalindra De Silva,
Nathan Gilbert, Ruihong Huang
School Of Computing
University of Utah
Salt Lake City, UT 84112
{riloff,asheq,alnds,ngilbert,huangrh}@cs.utah.edu, prafulla.surve@gmail.com

(a) Oh how I love *being ignored*. #sarcasm

(b) Thoroughly enjoyed *shoveling the driveway* today! :) #sarcasm

(c) Absolutely adore it when *my bus is late* #sarcasm

(d) I'm so pleased mom *woke me up* with vacuuming my room this morning. :) #sarcasm

Identified positive sentiment phrases and negative situation phrases using part of speech tags to build a sarcasm recognizer

Parsing

- Beyond annotating every word with a category or mapping spans to categories, more linguistic structure is enforced by how spans are combined in a sequence

Grammar

S → NP VP

VP → V NP|VP PP

PP → IN NP

V → eat

IN → with

NP → NP PP|we|sushi|chopsticks

These are called the production rules. Infinite sequences can be generated by following these rules

Syntax Tree

S → NP VP

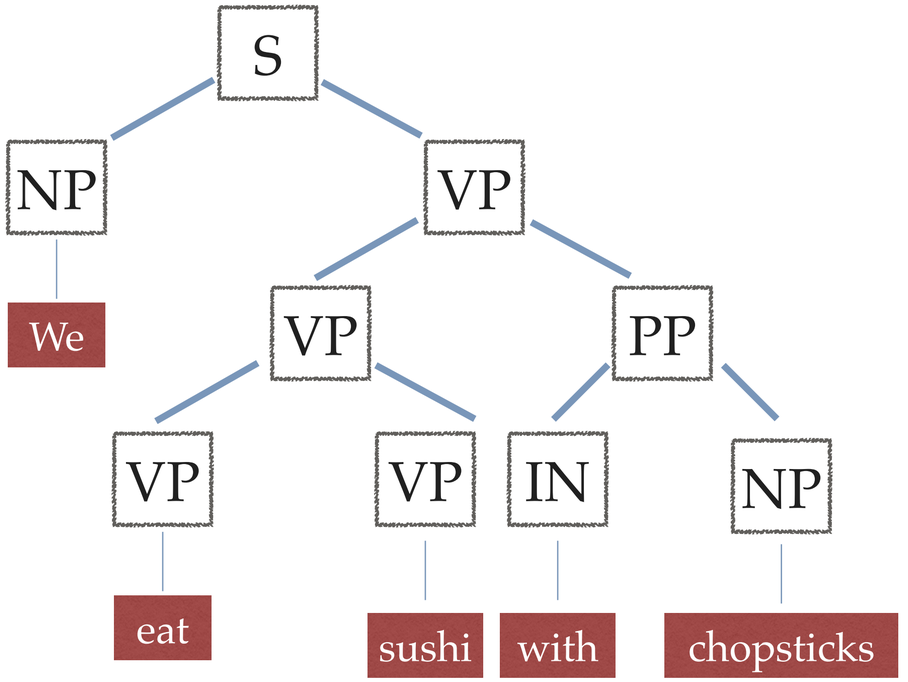
VP → V NP | VP PP

PP → IN NP

V → eat

IN → with

NP → NP PP | we | sushi | chopsticks



Parsing

S → NP VP

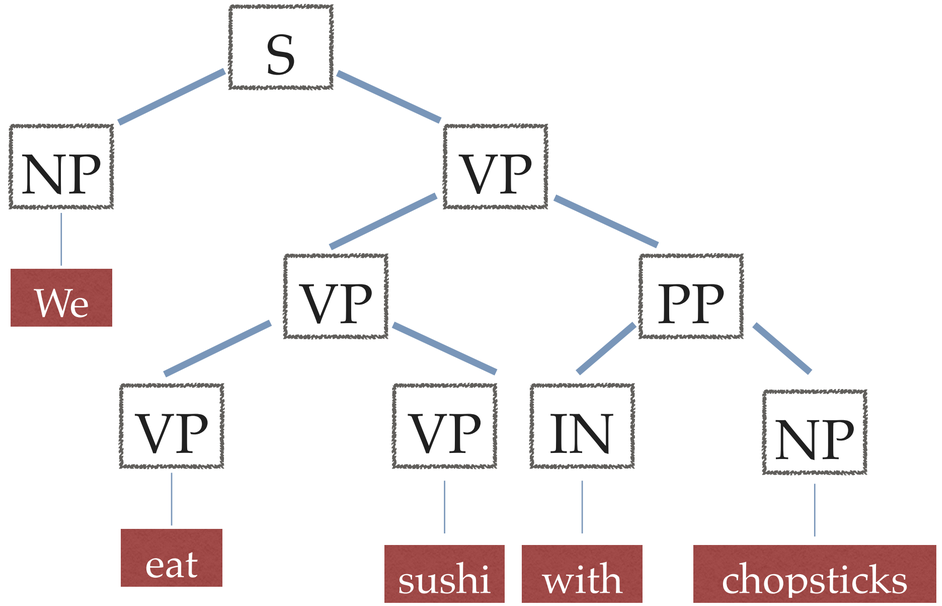
VP → V NP|VP PP

PP → IN NP

V → eat

IN → with

NP → NP PP|we|sushi|chopsticks

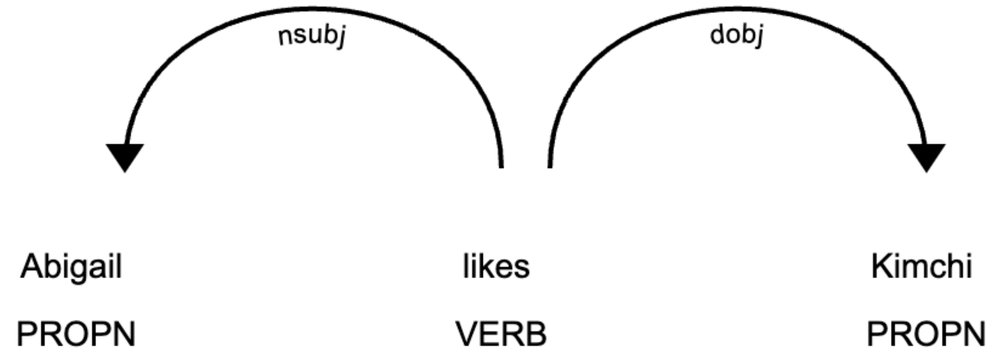


The goal of parsing is to learn the production rules and infer the most likely parse tree for a given sequence

Dependency parse

- Instead of decomposing a sequence into constituent subtrees, one can markup the sequence by labeling:
  - Important constituents (heads)
  - Their modifiers (dependents)
  - The type of relationship between the heads and dependents

Example



- What is happening? Something is being liked
- Who is liking? Abigail (the subject)
- What is being liked? Kimchi (the object)

Core dependents of clausal predicates

Nominal dep	*Predicate dep*	
nsubj	csubj	
↳ nsubj:pass	↳ csubj:pass	
↳ nsubj:outer	↳ csubj:outer	
obj	ccomp	xcomp
iobj		

Non-core dependents of clausal predicates

Nominal dep	*Predicate dep*	*Modifier word*
obl	advcl	advmod
↳ obl:npmod	↳ advcl:relcl	
↳ obl:tmod		

Special clausal dependents

Nominal dep	*Auxiliary*	*Other*
vocative	aux	mark
discourse	↳ aux:pass	
expl	cop	

Noun dependents

Nominal dep	*Predicate dep*	*Modifier word*
nummod	acl	amod
	↳ acl:relcl	
appos		det
		↳ det:predet
nmod		
↳ nmod:npmod		
↳ nmod:tmod		
↳ nmod:poss		

Compounding and unanalyzed

compound	flat
↳ compound:prt	↳ flat:foreign
fixed	goeswith

Coordination

conj	cc
	↳ cc:preconj

Case-marking, prepositions, possessive

case

Loose joining relations

list	parataxis	orphan
dislocated		reparandum

Other

Sentence head	*Punctuation*	*Unspecified dependency*
root	punct	dep

There are many syntactic dependencies that can be identified (example taken from universal dependencies treebank)

What can we do with dependency parse?

"Mistakes were made."

by whom?
by you?
by us?!

Source: NYTimes

Drivers of English Syntactic Change in the Canadian Parliament

Liwen Hou
Khoury College of Computer Sciences
Northeastern University
Boston, MA
hou.l@northeastern.edu

David A. Smith
Khoury College of Computer Sciences
Northeastern University
Boston, MA
dasmith@ccs.neu.edu

Abstract

Corpus linguists have long noted the "colloquialization" of many genres of English. While the average decline in many features of formal speech is obvious in aggregate, we are better able to disentangle drivers of change by examining Canadian parliamentary speeches coded for characteristics of individual speakers across more than 100 years—much longer than previous studies of individuals' language change in a common environment. While many language changes proceed by cohort replacement and often originate with female speakers, the Canadian Hansard shows that

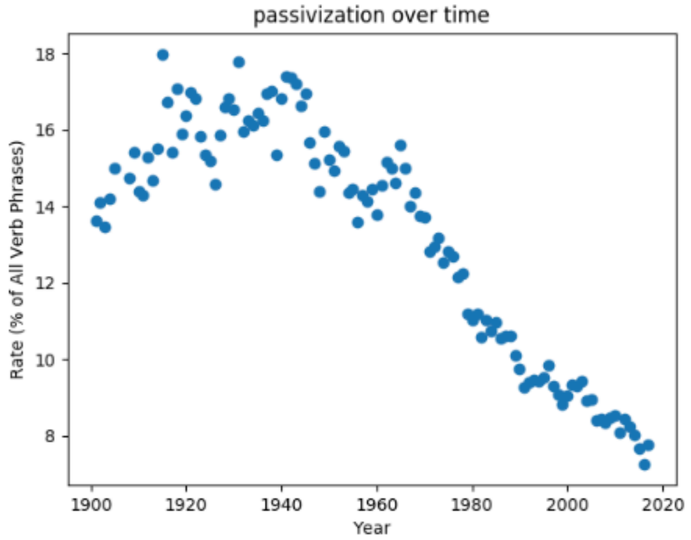
type of language change, we focus on syntactic aspects of colloquialization in this work.

Specifically, we find that the *be*-passive became rarer, that the progressive became more common, that pied-piping became less frequent, that modals such as "shall" also decreased in frequency, and that semi-modals such as "have to" rose in frequency. See Section 3 (and in particular Table 1) for examples of the five aforementioned features.

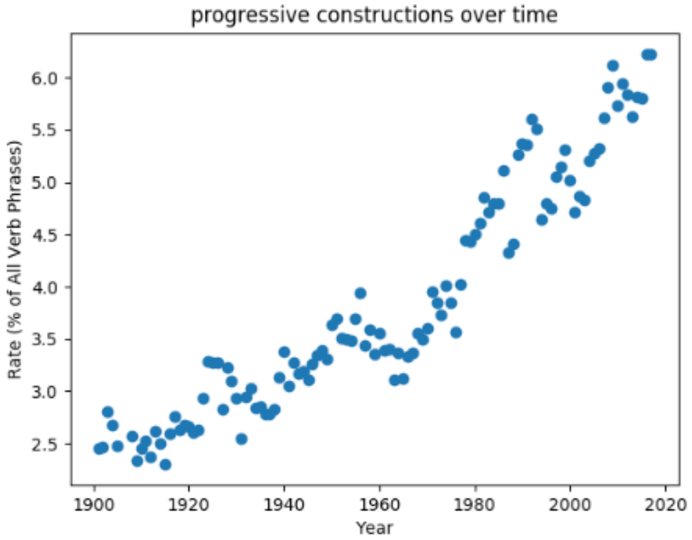
We are particularly interested in the forces driving language change. One natural hypothesis is that newer parliamentary members repeatedly introduce more colloquial language in their speeches, so that

Syntactic changes

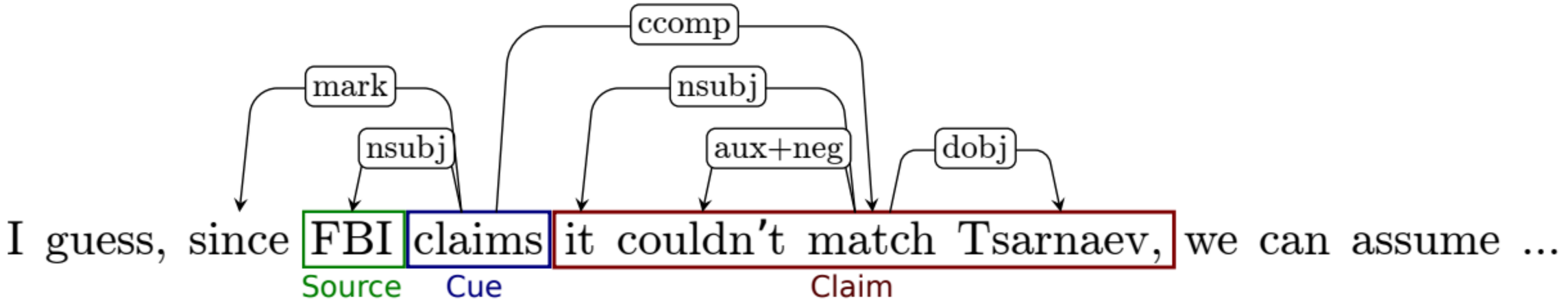
- Colloquialization increasing in political language
- Governing party less colloquial
- Women leaders less colloquial

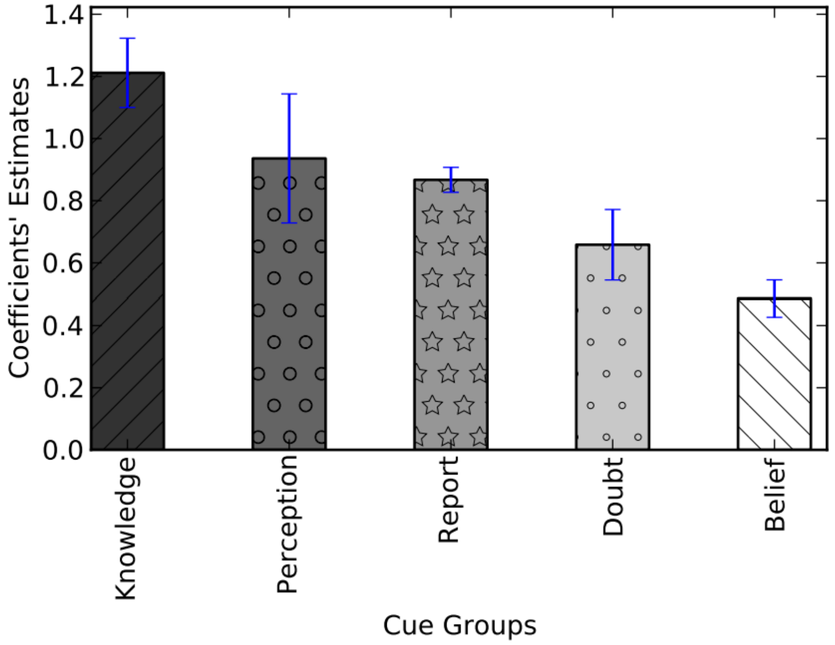


(a) Passivization



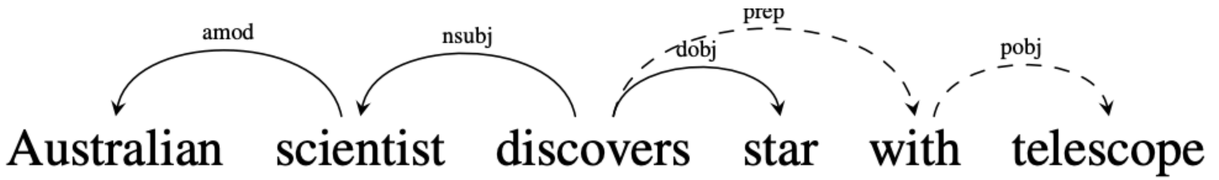
(b) Progressives

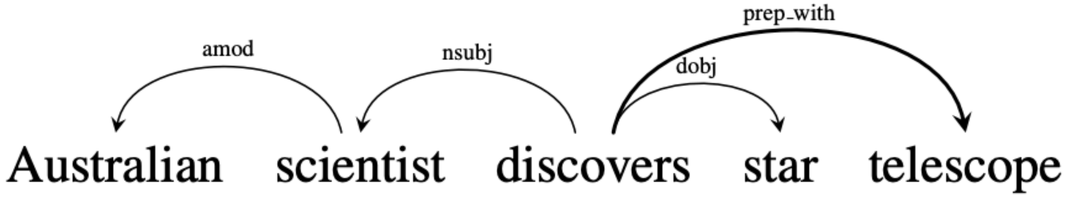




Using the dependency parse, we can identify the different parts of a sentence

- Sandeep Soni, Tanushree Mitra, Eric Gilbert, and Jacob Eisenstein. 2014. Modeling Factuality Judgments in Social Media Text. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 415–420, Baltimore, Maryland. Association for Computational Linguistics.





WORD	CONTEXTS
australian	scientist/amod^{-1}
scientist	australian/amod, discovers/nsubj^{-1}
discovers	scientist/nsubj, star/dobj, telescope/prep_with
star	discovers/dobj^{-1}
telescope	discovers/prep_with^{-1}

florida	gainesville	fla	texas
	fla	alabama	louisiana
	jacksonville	gainesville	georgia
	tampa	tallahassee	california
	lauderdale	texas	carolina

Similar words (last column is based on dependent parsing contexts)

Omer Levy and Yoav Goldberg. 2014. Dependency-Based Word Embeddings. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 302–308, Baltimore, Maryland. Association for Computational Linguistics.

In class

- Sequence labeling demo