ITCS 6101/8101: Natural Language Processing

Introduction

Razvan C. Bunescu

Department of Computer Science @ CCI

razvan.bunescu@charlotte.edu

What is Natural Language Processing?

- **Natural Language Processing (NLP)** = developing computer systems that can *process*, *understand*, or *communicate* in natural language (*text* or *speech*):
  - **Natural Languages**: English, Turkish, Japanese, Latin, Hawaiian Creole, Esperanto, American Sign Language, …
    - Music?
  - **Formal Languages**: C++, Java, Python, XML, OWL, Predicate Calculus, Lambda Calculus, …
  - **Natural Languages are significantly more difficult to process than Artificial Languages!**

- Major NLP tasks:
  - *machine translation*, *information extraction*, *question answering*, *taking instructions*, *holding conversations, code generation, …*

NLP Application: Question Answering

- Input:
  - A **question** + a large collection of **text documents**.
- Output:
  - An **answer**, or list of answers, found by '*mining*' the documents in the collection.

Who is the author of "Progress and Poverty"?

> *Progress and Poverty is an 1879 treatise on the questions of why poverty accompanies economic and technological progress and why economies exhibit a tendency toward cyclical boom and bust. It was written by the social theorist and economist Henry George, who is also variously known for leading an early movement that popularized Universal Basic Income, sporting a fancy beard while shouting "The Rent Is Too Damn High!" and inspiring a popular board game that was shamelessly ripped off and repackaged as Monopoly.*

NLP Application: Question Answering

Who is the author of “Progress and Poverty”?

> *Progress and Poverty is an 1879 treatise on the questions of why poverty accompanies economic and technological progress and why economies exhibit a tendency toward cyclical boom and bust. It was written by the social theorist and economist Henry George, who is also variously known for leading an early movement that popularized Universal Basic Income, sporting a fancy beard while shouting "The Rent Is Too Damn High!" and inspiring a popular board game that was shamelessly ripped off and repackaged as Monopoly.*

Try simple **string matching**:

The author of “Progress and Poverty” is …

NLP Application: Question Answering

Who is the author of "**Progress and Poverty**"?

> ***Progress and Poverty*** *is an 1879 treatise on the questions of why poverty accompanies economic and technological progress and why economies exhibit a tendency toward cyclical boom and bust. It was written by the social theorist and economist Henry George, who is also variously known for leading an early movement that popularized Universal Basic Income, sporting a fancy beard while shouting "The Rent Is Too Damn High!" and inspiring a popular board game that was shamelessly ripped off and repackaged as Monopoly.*

Tokenization:

{*Progress*, *and*, *Poverty*, *"*, *?*, *is*, *an*, *1879*, …, *George*, …, *as*, *Monopoly*, *.*} are tokens.

NLP Application: Question Answering

Who is the author of "**Progress and Poverty**"?

> ***Progress and Poverty*** *is an 1879 treatise on the questions of why poverty accompanies economic and technological progress and why economies exhibit a tendency toward cyclical boom and bust. It was written by the social theorist and economist Henry George, who is also variously known for leading an early movement that popularized Universal Basic Income, sporting a fancy beard while shouting "The Rent Is Too Damn High!" and inspiring a popular board game that was shamelessly ripped off and repackaged as Monopoly.*

NLP Application: Question Answering

Who is the author of "**Progress and Poverty**"?

Progress and Poverty *is an 1879* ***treatise*** *on the questions of why poverty accompanies economic and technological progress and why economies exhibit a tendency toward cyclical boom and bust.* ***It*** *was written by the social theorist and economist Henry George, who is also variously known for leading an early movement that popularized Universal Basic Income, sporting a fancy beard while shouting "The Rent Is Too Damn High!" and inspiring a popular board game that was shamelessly ripped off and repackaged as Monopoly.*

Coreference resolution:

{*Progress and Poverty*, *treatise*, *It*} are coreferent.

NLP Application: Question Answering

Who is the author of "**Progress and Poverty**"?

> ***Progress and Poverty*** *is an 1879* ***treatise*** *on the questions of why poverty accompanies economic and technological progress and why economies exhibit a tendency toward cyclical boom and bust.* ***It*** *was written by the social theorist and economist Henry George, who is also variously known for leading an early movement that popularized Universal Basic Income, sporting a fancy beard while shouting "The Rent Is Too Damn High!" and inspiring a popular board game that was shamelessly ripped off and repackaged as Monopoly.*

NLP Application: Question Answering

Who is the **author of** "**Progress and Poverty**"?

Progress and Poverty *is an 1879* ***treatise*** *on the questions of why poverty accompanies economic and technological progress and why economies exhibit a tendency toward cyclical boom and bust.* ***It*** *was* ***written by*** *the social theorist and economist* ***Henry George****, who is also variously known for leading an early movement that popularized Universal Basic Income, sporting a fancy beard while shouting "The Rent Is Too Damn High!" and inspiring a popular board game that was shamelessly ripped off and repackaged as Monopoly.*

Syntactic parsing:

- *It* is the *nominal* ***subject*** of *written*; *George* is the *agent* ***argument*** of *written*.

NLP Application: Question Answering

Who is the **author of** "**Progress and Poverty**"?

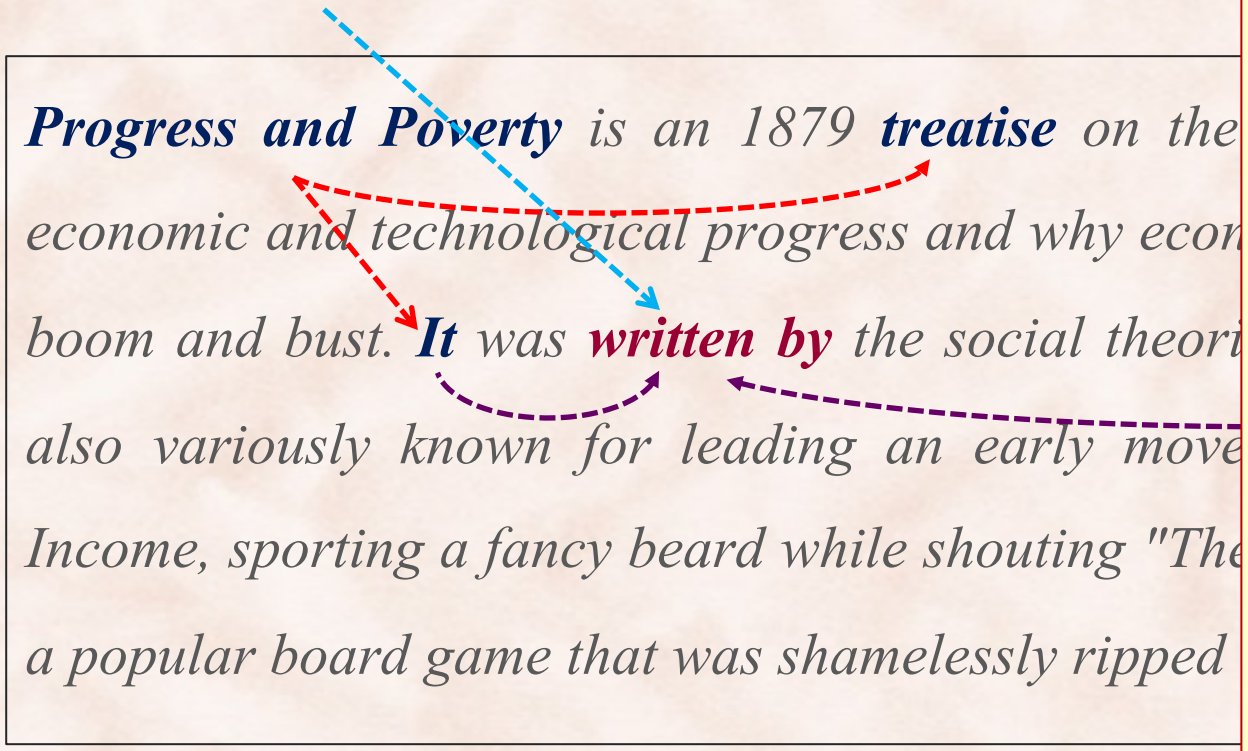
Progress and Poverty *is an 1879* ***treatise*** *on the questions of why poverty accompanies economic and technological progress and why economies exhibit a tendency toward cyclical boom and bust.* ***It*** *was* ***written by*** *the social theorist and economist* ***Henry George****, who is also variously known for leading an early movement that popularized Universal Basic Income, sporting a fancy beard while shouting "The Rent Is Too Damn High!" and inspiring a popular board game that was shamelessly ripped off and repackaged as Monopoly.*

Logical inference or **reasoning**:

- X was *written by* Y => Y is the *author of* X
  - *Prog and Pov* was *written by George* => *George* is the *author* of *Prog and Pov*

NLP Application: Question Answering

Who is the **author of** "**Progress and Poverty**"?



Fundamental NLP tasks

- **Tokenization**.
- **Syntactic Analysis**:
  - Syntactic Dependencies.
- **Word Meaning** representations.
- **Coreference Resolution**.
- **Semantic Parsing**.
  - Logical Forms.

Fundamental AI tasks

- Knowledge Representation.
  - First Order Predicate Logic.
    - **Inference** in FOPL.

Fundamental NLP Tasks Identify **Linguistic Structures**

- Tokenization
- Morphological Analysis
- Part of Speech Tagging
- Syntactic Parsing
- Word Meaning
- Semantic Parsing
- Anaphora/Coreference Resolution
- Named Entity Linking

} Morphology

} Syntax

} Semantics

} Discourse and Pragmatics

Tokenization

Words, Morphemes, Subwords

Part of Speech (POS) Tagging

- Annotate each word in a sentence with its POS:
  - nouns, verbs, adjectives, adverbs, pronouns, prepositions, …

PRP VBD TO VB TO DT NN IN NN VBD VBG

They used to object to the use of object oriented programming

obJECT OBject

- Useful for many NLP tasks downstream:
  - speech recognition and synthesis, syntactic parsing, word sense disambiguation, information retrieval, …

- Superseeded in many tasks by (contextualized) ***word embeddings***.

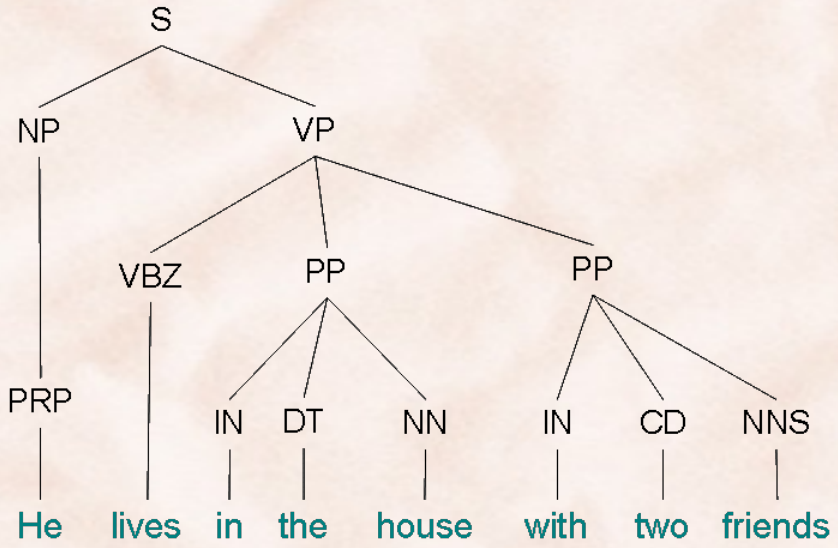
Core 36 POS Tags from Penn Treebank

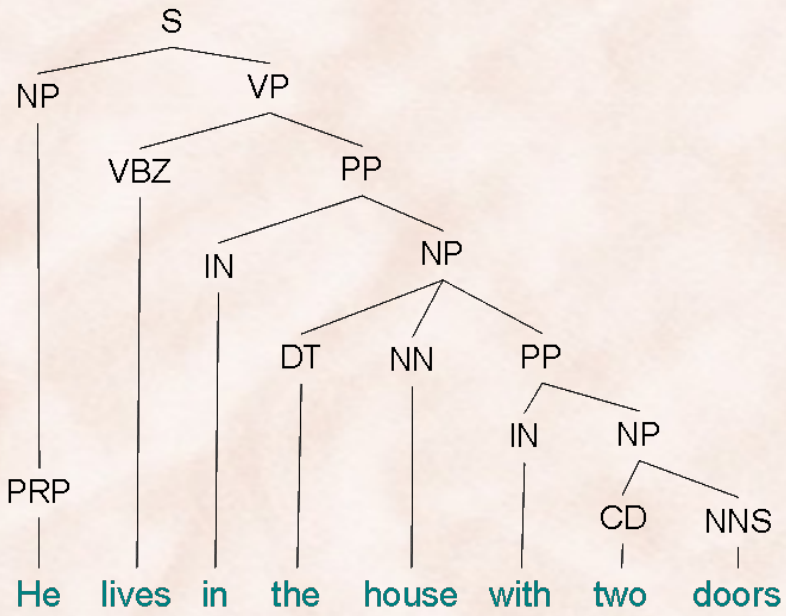
Tag	Description	Example	Tag	Description	Example	Tag	Description	Example
CC	coord. conj.	*and, but, or*	NNP	proper noun, sing.	*IBM*	TO	infinitive to	*to*
CD	cardinal number	*one, two*	NNPS	proper noun, plu.	*Carolinas*	UH	interjection	*ah, oops*
DT	determiner	*a, the*	NNS	noun, plural	*llamas*	VB	verb base	*eat*
EX	existential 'there'	*there*	PDT	predeterminer	*all, both*	VBD	verb past tense	*ate*
FW	foreign word	*mea culpa*	POS	possessive ending	*'s*	VBG	verb gerund	*eating*
IN	preposition/ subordin-conj	*of, in, by*	PRP	personal pronoun	*I, you, he*	VBN	verb past participle	*eaten*
JJ	adjective	*yellow*	PRP$	possess. pronoun	*your*	VBP	verb non-3sg-pr	*eat*
JJR	comparative adj	*bigger*	RB	adverb	*quickly*	VBZ	verb 3sg pres	*eats*
JJS	superlative adj	*wildest*	RBR	comparative adv	*faster*	WDT	wh-determ.	*which, that*
LS	list item marker	*1, 2, One*	RBS	superlatv. adv	*fastest*	WP	wh-pronoun	*what, who*
MD	modal	*can, should*	RP	particle	*up, off*	WP$	wh-possess.	*whose*
NN	sing or mass noun	*llama*	SYM	symbol	+, %, &	WRB	wh-adverb	*how, where*

Figure 17.2 Penn Treebank core 36 part-of-speech tags.

Syntax: Constituency Parsing

- Compute the *phrase structure* of a sentence:





Parse Tree example for simple grammar

Grammar Rules			Examples
S	→	*NP VP*	I + want a morning flight
NP	→	*Pronoun*	I
	|	*Proper-Noun*	Los Angeles
	|	*Det Nominal*	a + flight
Nominal	→	*Nominal Noun*	morning + flight
	|	*Noun*	flights
VP	→	*Verb*	do
	|	*Verb NP*	want + a flight
	|	*Verb NP PP*	leave + Boston + in the morning
	|	*Verb PP*	leaving + on Thursday
PP	→	*Preposition NP*	from + Los Angeles

Figure 18.3 The grammar for $\mathscr{L}_0$, with example phrases for each rule.

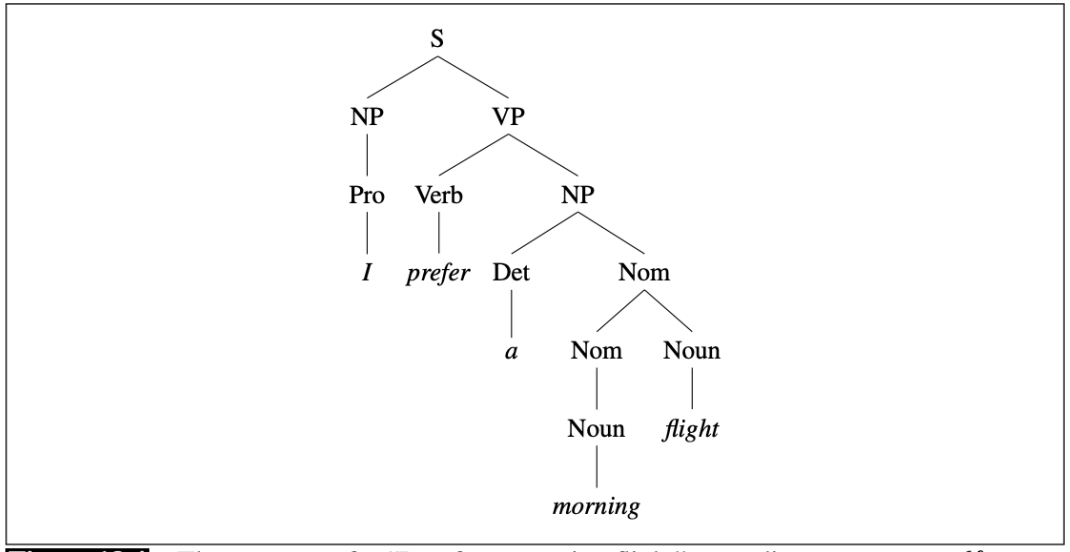


Figure 18.4 The parse tree for "I prefer a morning flight" according to grammar $\mathscr{L}_0$.

Parse Tree example from Treebank corpus

```
((S
   (NP-SBJ (DT That)
     (JJ cold) (, ,)
     (JJ empty) (NN sky) )
   (VP (VBD was)
     (ADJP-PRD (JJ full)
       (PP (IN of)
         (NP (NN fire)
           (CC and)
           (NN light) ))))
   (. .) ))
```

(a)

```
((S
   (NP-SBJ The/DT flight/NN )
   (VP should/MD
     (VP arrive/VB
       (PP-TMP at/IN
         (NP eleven/CD a.m/RB ))
       (NP-TMP tomorrow/NN )))))
```

(b)

Figure 18.5 Parses from the LDC Treebank3 for (a) Brown and (b) ATIS sentences.

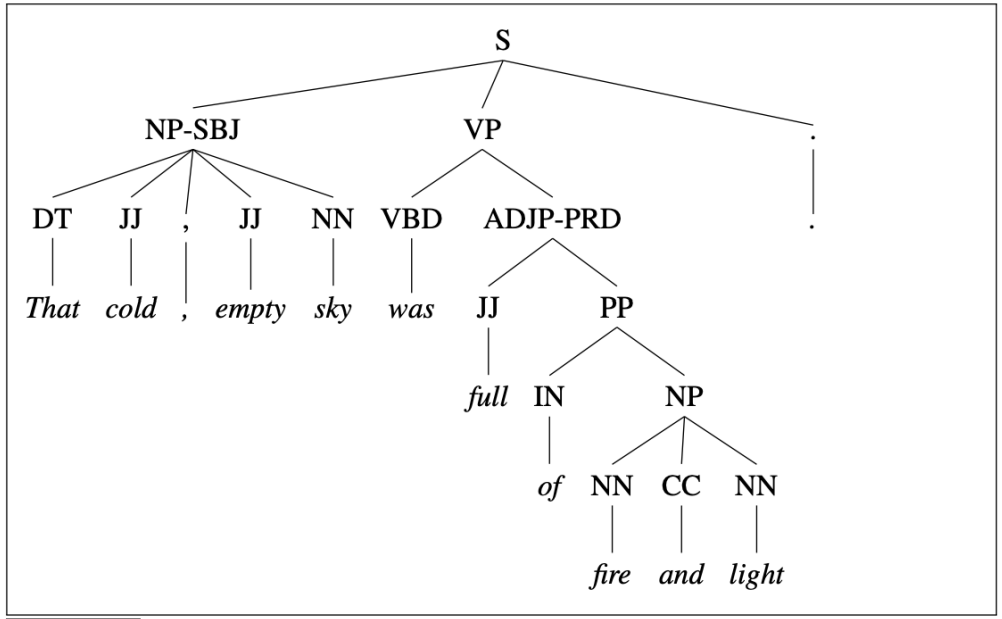
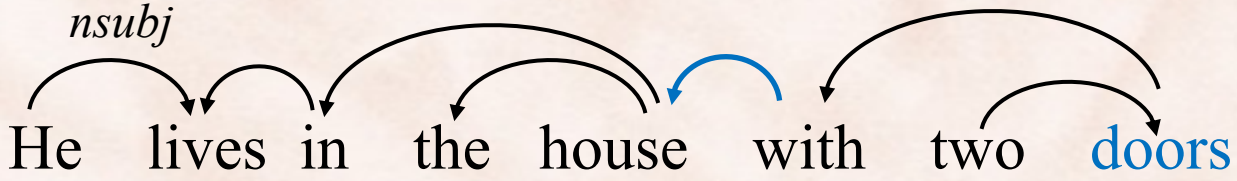


Figure 18.6 The tree corresponding to the Brown corpus sentence in the previous figure.

Syntax: Dependency Parsing



Universal Dependency Relations

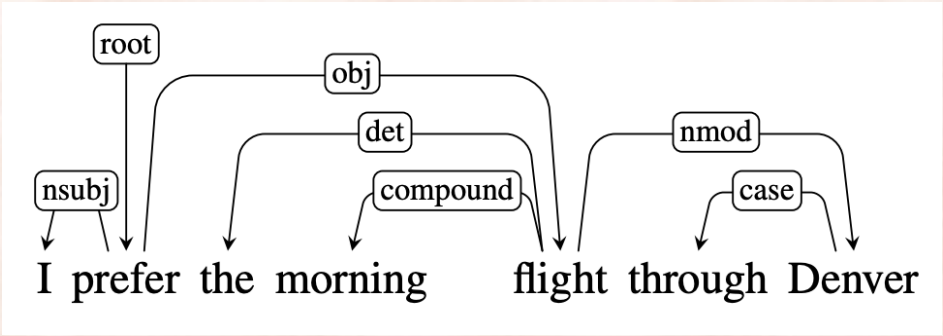
Clausal Argument Relations	Description
NSUBJ	Nominal subject
OBJ	Direct object
IOBJ	Indirect object
CCOMP	Clausal complement
Nominal Modifier Relations	**Description**
NMOD	Nominal modifier
AMOD	Adjectival modifier
APPOS	Appositional modifier
DET	Determiner
CASE	Prepositions, postpositions and other case markers
Other Notable Relations	**Description**
CONJ	Conjunct
CC	Coordinating conjunction

Figure 19.2 Some of the Universal Dependency relations (de Marneffe et al., 2021).

Dependency vs. Constituency Structures

Dependency trees can be automatically extracted from phrase structure trees:

- Generating Typed Dependency Parses from Phrase Structure Parses, de Marneffe et al., LREC 2006.



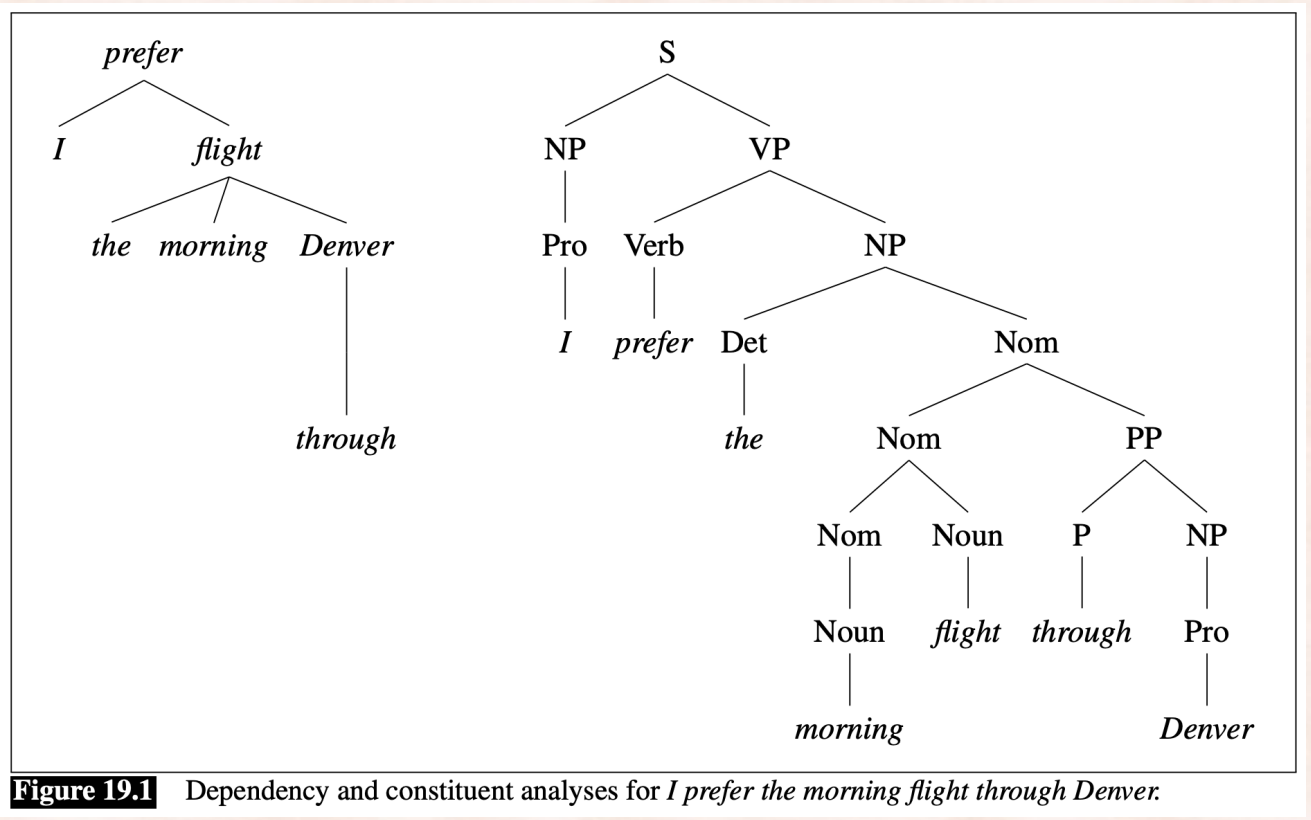
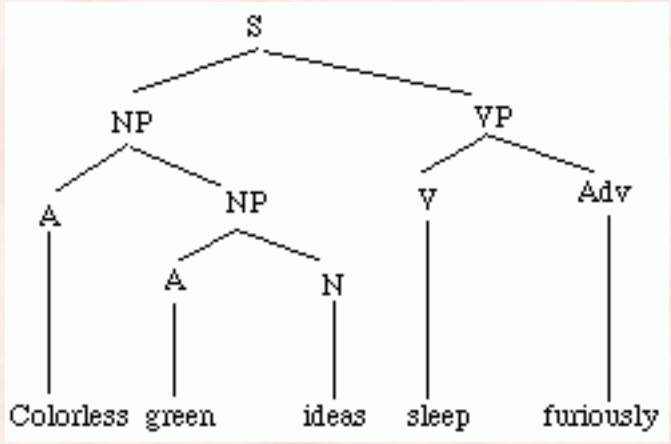


Figure 19.1 Dependency and constituent analyses for *I prefer the morning flight through Denver.*

Syntax vs. Semantics

- **The to runs the cat dog after**.
  - Syntactically incorrect.
    - Cannot build phrase structure (parse tree).
  - Semantically meaningless.
    - Cannot have meaning if sentence is not well formed syntactically.

- **Colorless green ideas sleep furiously**:
  - Syntactically correct.
  - Semantically meaningless (literal reading).



- **The fat cat sat on the mat**:
  - Syntactically correct.
  - Semantically meaningful: $cat(C) \wedge mat(M) \wedge satOn(C, M)$.
  - Pragmaticaly?

Colorless green ideas sleep furiously: Figurative Reading

It can only be the thought of verdure to come, which prompts us in the autumn to buy these dormant white lumps of vegetable matter covered by a brown papery skin, and lovingly to plant them and care for them. It is a marvel to me that under this cover they are labouring unseen at such a rate within to give us the sudden awesome beauty of spring flowering bulbs. While winter reigns the earth reposes but these ***colourless green ideas sleep furiously***.

[by C. M. Street]

http://linguistlist.org/issues/2/2-457.html#2

Word Meaning

- Words in natural language may have multiple meanings:
  - he cashed a check at the bank
  - he sat on the bank of the river and watched the currents
  - they built a large plant to manufacture automobiles
  - chlorophyll is generally present in plant leaves
- Use lexical resources such as WordNet that map words to their ***discrete meanings***:
  - http://wordnetweb.princeton.edu/perl/webwn
- Identifying the meaning of a word is useful for:
  - machine translation, information retrieval, question answering, text classification, …
- Nowadays superseded in many tasks by (contextualized) ***word embeddings***.

Logical Forms (Semantic Parsing)

- Map natural language sentences to a formal semantic representation (*logic form*).
- Text to SQL, for interaction with DBs in natural language:
  - *List all song names by singers age above the average singer age.*
  - SELECT song_name FROM singers WHERE age > (SELECT avg(age) FROM singers)
- In RoboCup, map coaching advice to Clang:
  - *If the ball is in our penalty area, all our players except player 4 should stay in our half.*
  - ((bpos (penalty-area our)) (do (player-except our {4}) (pos (half our))))
- In GeoQuery, map sentences to Prolog queries:
  - *How many states does the Mississippi run through?*
  - answer(A, count(B, (state(B), const(C, riverid(mississippi)), traverse(C, B)), A))

Logic Forms (Semantic Parsing)

- Automatic generation of code, e.g. for cards in Trading Card Games (TCGs):


```
name: [
  'D', 'i', 'r', 'e', ' ',
  'W', 'o', 'l', 'f', ' ',
  'A', 'l', 'p', 'h', 'a']
cost: ['2']
type: ['Minion']
rarity: ['Common']
race: ['Beast']
class: ['Neutral']
description: [
  'Adjacent', 'minions', 'have',
  '+', '1', 'Attack', '.']
health: ['2']
attack: ['2']
durability: ['-1']
```

```
class DireWolfAlpha(MinionCard):
  def __init__(self):
    super().__init__(
      "Dire Wolf Alpha", 2, CHARACTER_CLASS.ALL,
      CARD_RARITY.COMMON, minion_type=MINION_TYPE.BEAST)
  def create_minion(self, player):
    return Minion(2, 2, auras=[
      Aura(ChangeAttack(1), MinionSelector(Adjacent()))
    ])
```


```
class ManaWyrm(MinionCard):
  def __init__(self):
    super().__init__(
      'Mana Wyrm', 1,
      CHARACTER_CLASS.MAGE,
      CARD_RARITY.COMMON)
  def create_minion(self, player):
    return Minion(
      1, 3, effects=[
        Effect(
          SpellCast(),
          ActionTag(
            Give(ChangeAttack(1)),
                 SelfSelector()))
      ])
```

Coreference

- Determine which noun phrases refer to the same discourse entity.

Originally from Hawaii, Obama is a graduate of Columbia University and Harvard Law School, where he was the president of the Harvard Law Review. He was a community organizer in Chicago before earning his law degree.

Originally from Hawaii, Obama is a graduate of Columbia University and Harvard Law School, where he was the president of the Harvard Law Review. He was a community organizer in Chicago before earning his law degree.

The Curse of Ambiguity

- Computational Linguists are obsessed by ambiguity in NL:
  - unlike compiler writers.
- Ambiguity happens at all basic levels of natural language processing.
- Find at least 5 meanings of the following sentence:
  - I made her duck.

Ambiguity: “I made her duck”

1) I cooked waterfowl for her benefit (to eat).
2) I cooked waterfowl belonging to her.
3) I created the (plaster?) duck she owns.
4) I caused her to quickly lower her head or body.
5) I waved my magic wand and turned her into undifferentiated waterfowl.

Ambiguity: “I made her duck”

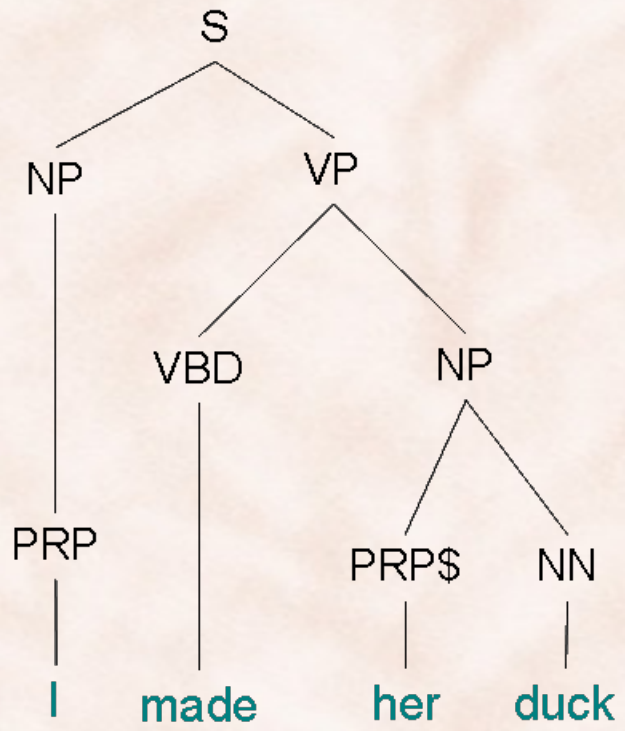
- **Part-of-Speech**: “duck” can be a N or V:
  - V: I caused her to quickly lower her head or body
  - N: I cooked waterfowl for her benefit (to eat).

- **Syntactic:** “her” can be a *possessive* (“of her”) or *dative* (“for her”) or *accusative* pronoun:
  - Possessive: I cooked waterfowl belonging to her.
  - Dative: I cooked waterfowl for her benefit (to eat).
  - Accusative: I waved my magic wand and turned her into waterfowl.

- **Semantic:** “make” can mean “create” or “cook”:
  - Create: I made the (plaster) duck statue she owns
  - Cook: I cooked waterfowl belonging to her.

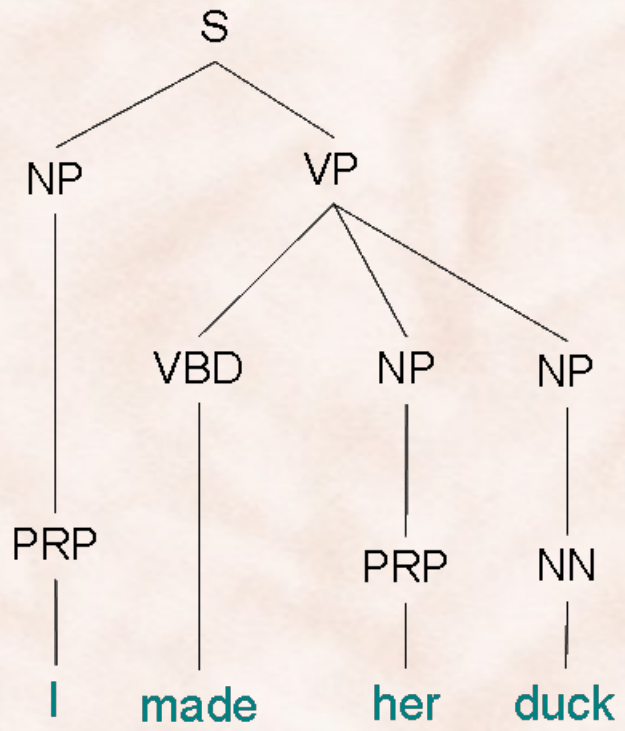
Ambiguity: “I made her duck”

- **Syntactic Parsing**:
  - **Make can be Transitive (verb has a noun direct object):**
    - I cooked [waterfowl belonging to her]



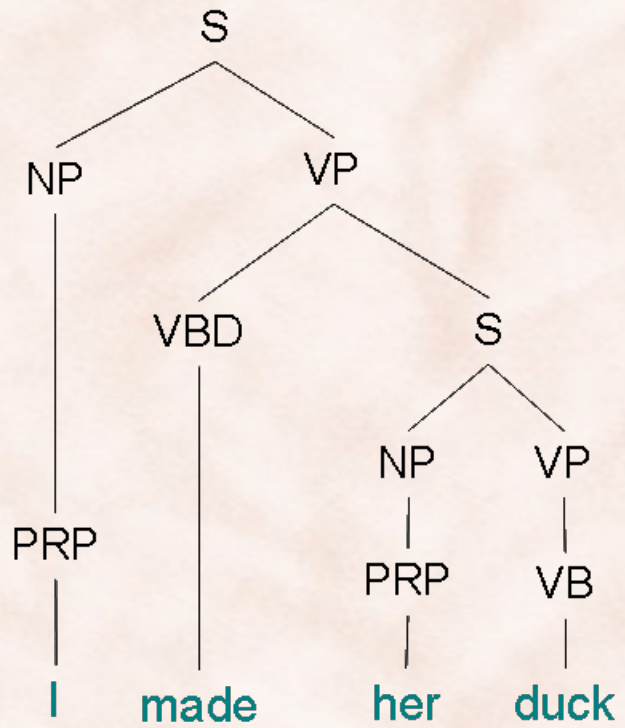
Ambiguity: “I made her duck”

- **Syntactic Parsing**:
  - **Make can be Ditransitive (verb has 2 noun objects):**
    - I made [her] (into) [undifferentiated waterfowl]



Ambiguity: “I made her duck”

- **Syntactic Parsing**:
  - **Make can be Action-transitive:**
    - I caused [her] [to move her body]



Ambiguity: “I made her duck”

- **Speech Recognition**:
  - I mate or duck
  - I’m eight or duck
  - Eye maid; her duck
  - Aye mate, her duck
  - I maid her duck
  - I’m aid her duck
  - I mate her duck
  - I’m ate her duck
  - I’m ate or duck

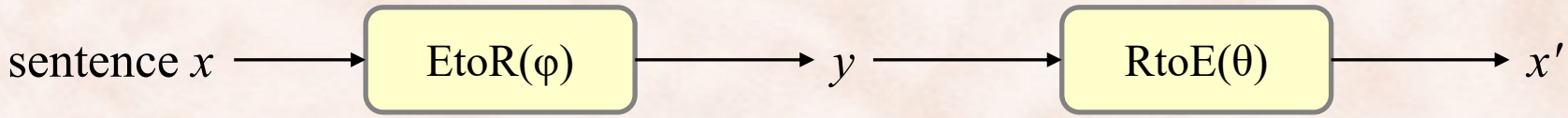
Ambiguity and Machine Translation

- English ⇒ Italian:
  - Mary **plays** the piano ⇒ Maria **suona** il pianoforte.
  - Mary **plays** with her cat ⇒ Maria **gioca** con il suo gatto.

- "Lost in translation" jokes from supposedly early MT system output (English ⇒ Russian ⇒ English):
  - "The spirit is willing, but the flesh is weak" => Russian => English:

    ⇒ The vodka is good, but the meat is spoiled.
  - "Out of sight, out of mind" => Russian => English:

    ⇒ Invisible idiot.
  - Modern MT systems use **backtranslation** as a constraint during ML training:



    - train parameters φ and θ such that $x \cong x'$

Ambiguity and Discourse: Coreference Resolution

- Winograd schemas:
  1. The **dog** chased the **cat**, which ran up a **tree**. **It** waited at the [top / bottom].
  2. The city **councilmen** refused the **demonstrators** a permit because **they** [feared / advocated] violence.
  3. The **trophy** doesn't fit into the brown **suitcase** because **it**'s too [small / large].
  4. **Paul** tried to call **George** on the phone, but **he** wasn't [successful / available].
  5. The **man** couldn't lift his **son** because **he** was so [weak / heavy].
  6. **Ann** asked **Mary** what time the library closes, [but / because] **she** had forgotten.

https://www.cs.nyu.edu/davise/papers/WS.html

Modality and Ambiguity: What does Nancy want?

- "**Nancy wants to mary an analytic philosopher**" [Eco, "Kant and the Platypus", 2000]

- Semantic interpretations:
  - [*de re*]: Nancy wants to marry a determined individual X, who is an analytic philosopher.

    $\exists x \Box Ax$
  - [*de dicto*]: Nancy wants to marry anybody, as long as he is an analytic philosopher.

    $\Box \exists x Ax$

- Pragmatic Interpretations (speaker's intentions):
  - Nancy wants to marry a determined individual, an analytic philosopher: she knows who he is, but the speaker doesn't, because she hasn't told him the name.
  - Nancy wants to marry a determined individual X, an analytic philosopher: she has also given the speaker the name and introduced them to each other, but out of discretion the speaker has thought it more fitting to avoid going into details.
  - …

Ambiguity is Pervasive in Natural Language

- CL and NLP are obsessed with ambiguity:
  - unlike compiler writers.
- Ambiguity happens at all basic levels of language processing.
- [Pros] Allows for significant compression of utterances:
  - people use context and knowledge about the world to disambiguate.
- [Cons] Very challenging for NLP!

Knowledge Involved in Resolving Ambiguity

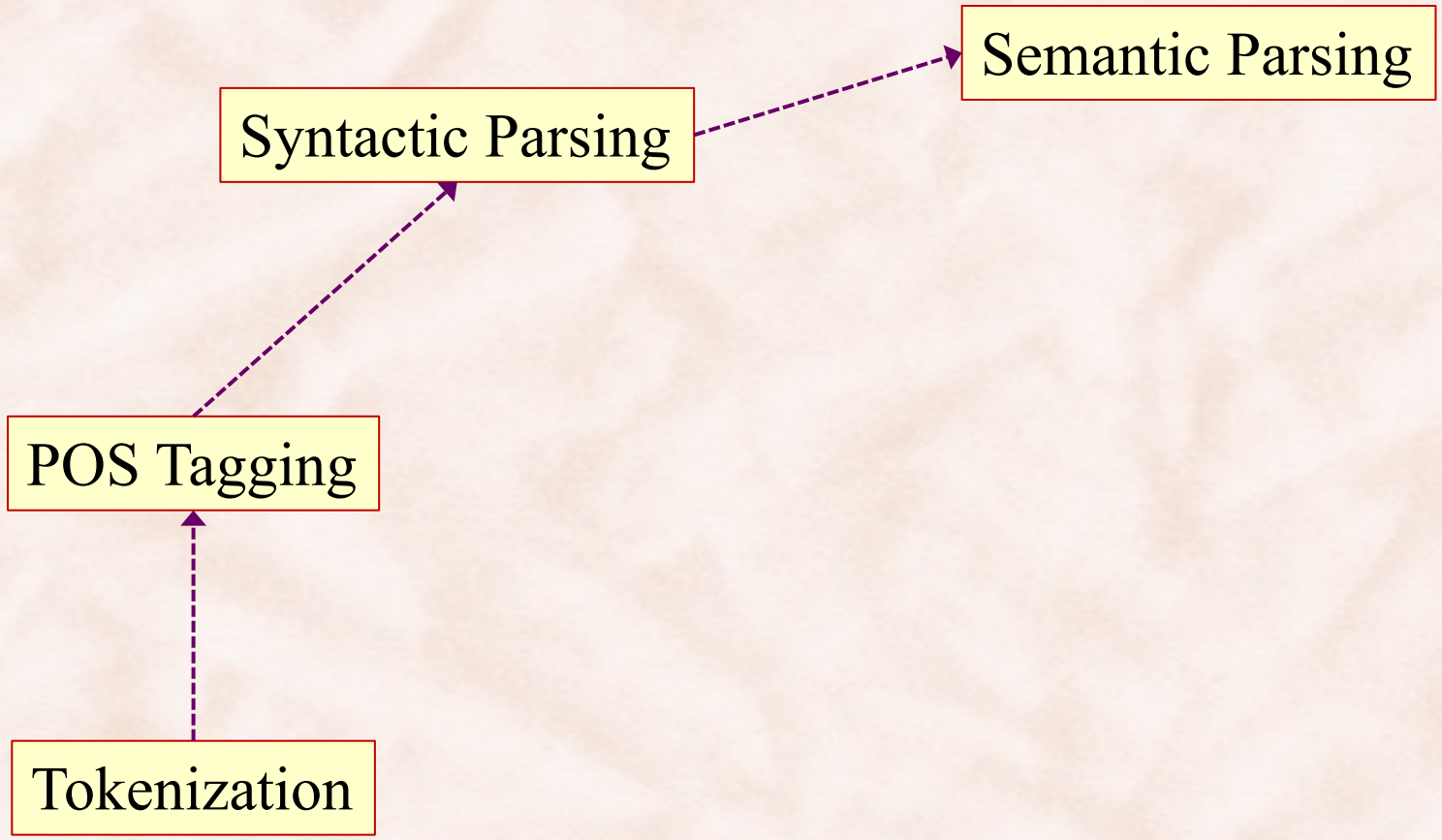
- **Syntax**:
  - An agent is typically the subject of the verb (SRL).
- **Semantics**:
  - John and Mary are names of people.
  - Columbus and Athens are city names.
- **Pragmatics**:
  - If she is hungry and she is not vegetarian, it is likely she will enjoy cooked duck.
- **Word knowledge**:
  - Houses have a (variable number of) doors.
  - An individual may live with other people (friends) in the same house.

Two Major Approaches to NLP

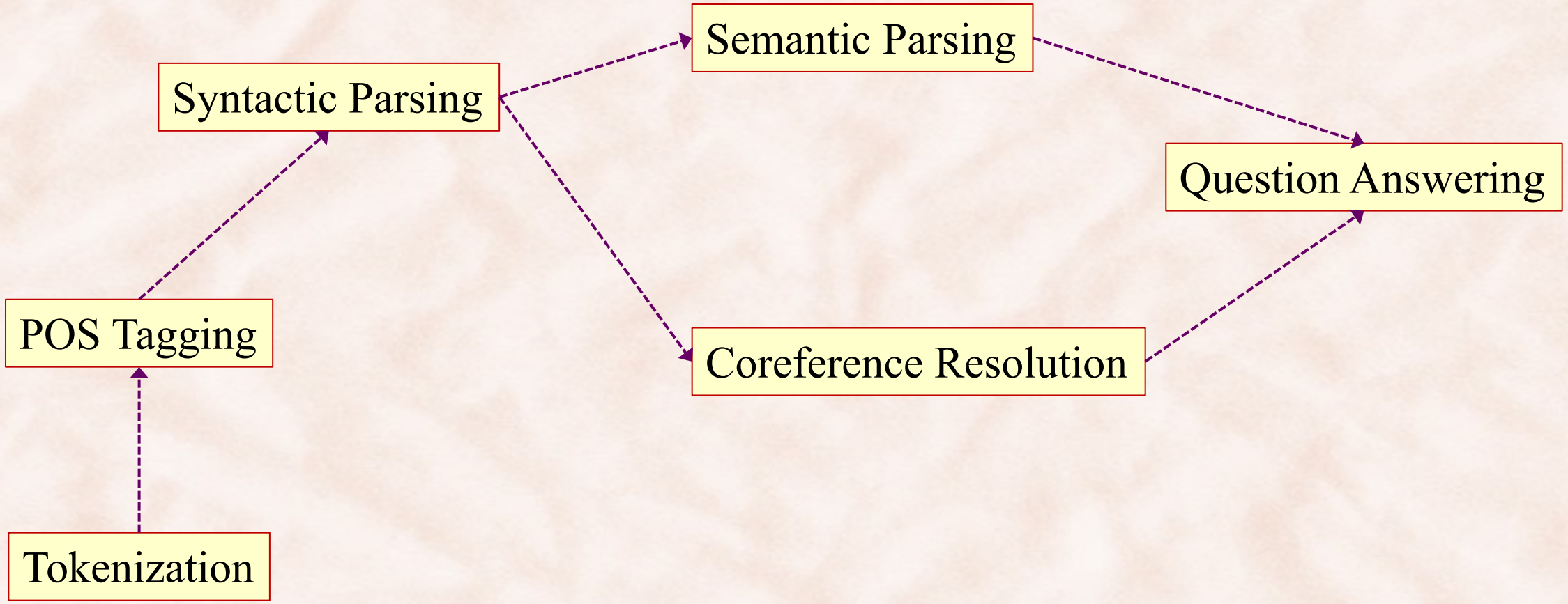
1. **NLP Pipelines**:
   - Transforming text into a stack of general-purpose linguistic structures:
     - From subword units called *morphemes*, to word-level *parts-of-speech*, to *tree-structured representations* of grammar, and beyond, to *logic-based representations* of meaning.
   - These general-purpose structures should then be able to support any desired application.
   - Should generalize better across applications, scenarios, and languages:
     - Linguistic structure seems to be particularly important when training data is limited.

2. **End-to-End** (*learning from scratch*):
   - Started by [Collobert et al., 2011], transform raw text into any desired output structure:
     - a summary, a database, a translation, …
   - **Language Models**: learning to predict words in context or just the next words seems sufficient for learning all levels of linguistic knowledge, as well as world knowledge.

NLP Pipeline Example



NLP Pipeline Example



End-to-End: Semantic Role Labeling

"*End-to-end Learning of Semantic Roles using RNNs* [Zhou & Xu, ACL 2015]

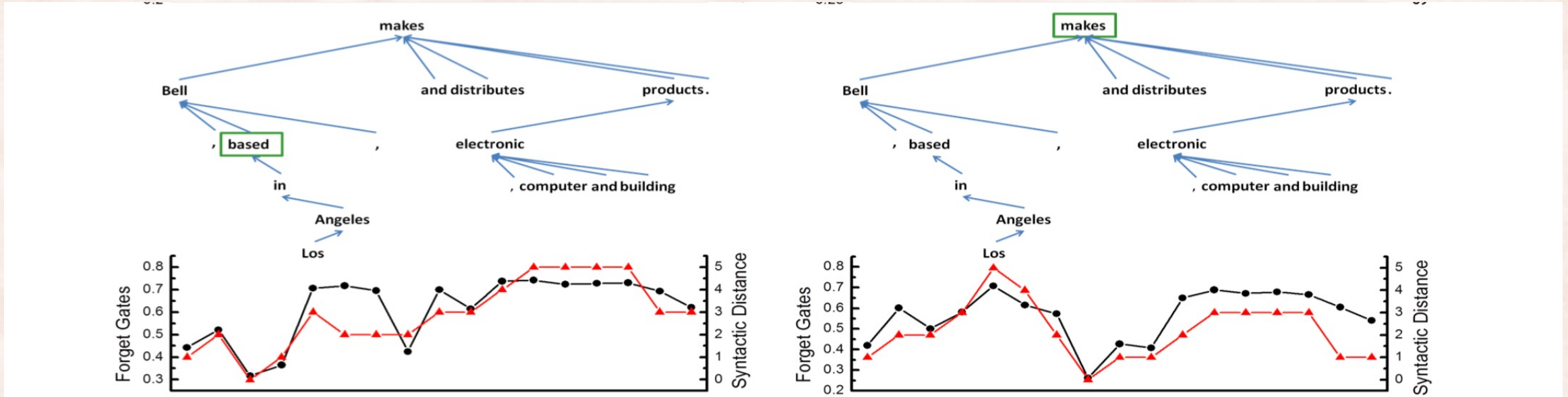
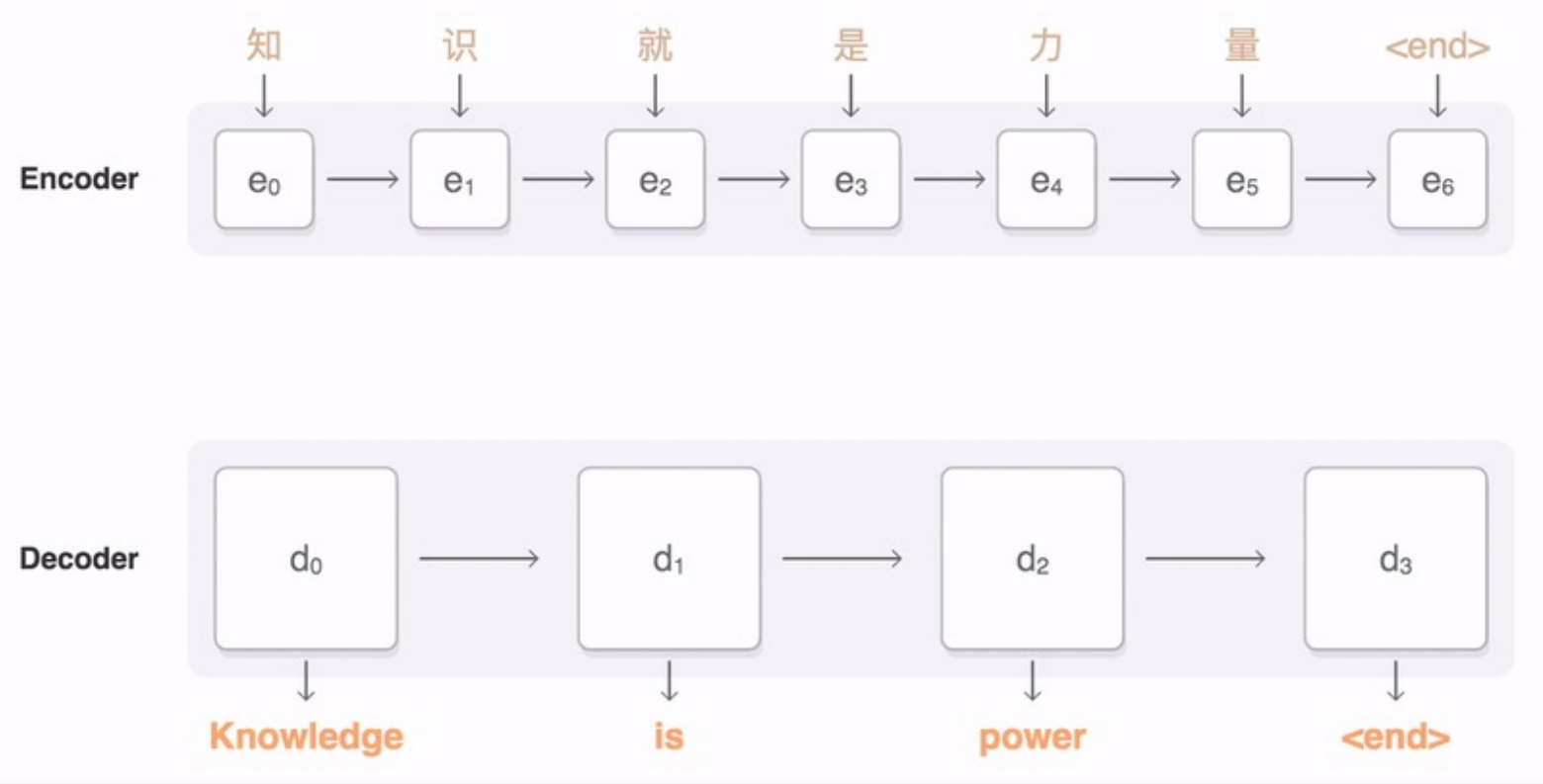


Figure 6: Forget gates value vs. Syntactic distance on four example sentences. Top: dependency parsing tree from gold tag. Green square word: predicate word. Bottom black solid lines: forget gates value at each time step. Bottom red empty square lines: gold syntactic distance between the current argument and predicate.

End-to-End: Machine Translation with RNNs & Attention

- From Phrase-Based Machine Translation to Neural Machine Translation:

  https://research.googleblog.com/2016/09/a-neural-network-for-machine.html



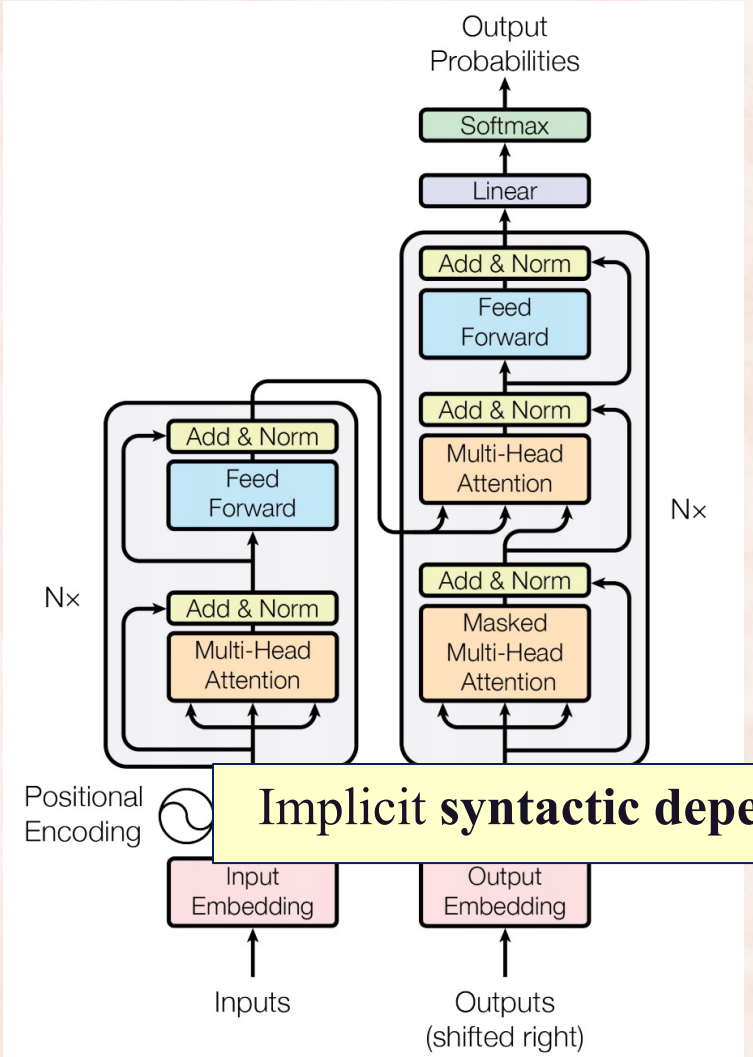
End-to-End: Machine Translation with RNNs & Attention

- Before November 2016:
  - Kilimanjaro is 19,710 feet of the mountain covered with snow, and it is said that the highest mountain in Africa. Top of the west, “Ngaje Ngai” in the Maasai language, has been referred to as the house of God. The top close to the west, there is a dry, frozen carcass of a leopard. Whether the leopard had what the demand at that altitude, there is no that nobody explained.
- After November 2016:
  - Kilimanjaro is a mountain of 19,710 feet covered with snow and is said to be the highest mountain in Africa. The summit of the west is called “Ngaje Ngai” in Masai, the house of God. Near the top of the west there is a dry and frozen dead body of leopard. No one has ever explained what leopard wanted at that altitude.

http://www.nytimes.com/2016/12/14/magazine/the-great-ai-awakening.html

End-to-End: Machine Translation With Transformer

https://arxiv.org/pdf/1706.03762.pdf



Implicit **syntactic dependences**.

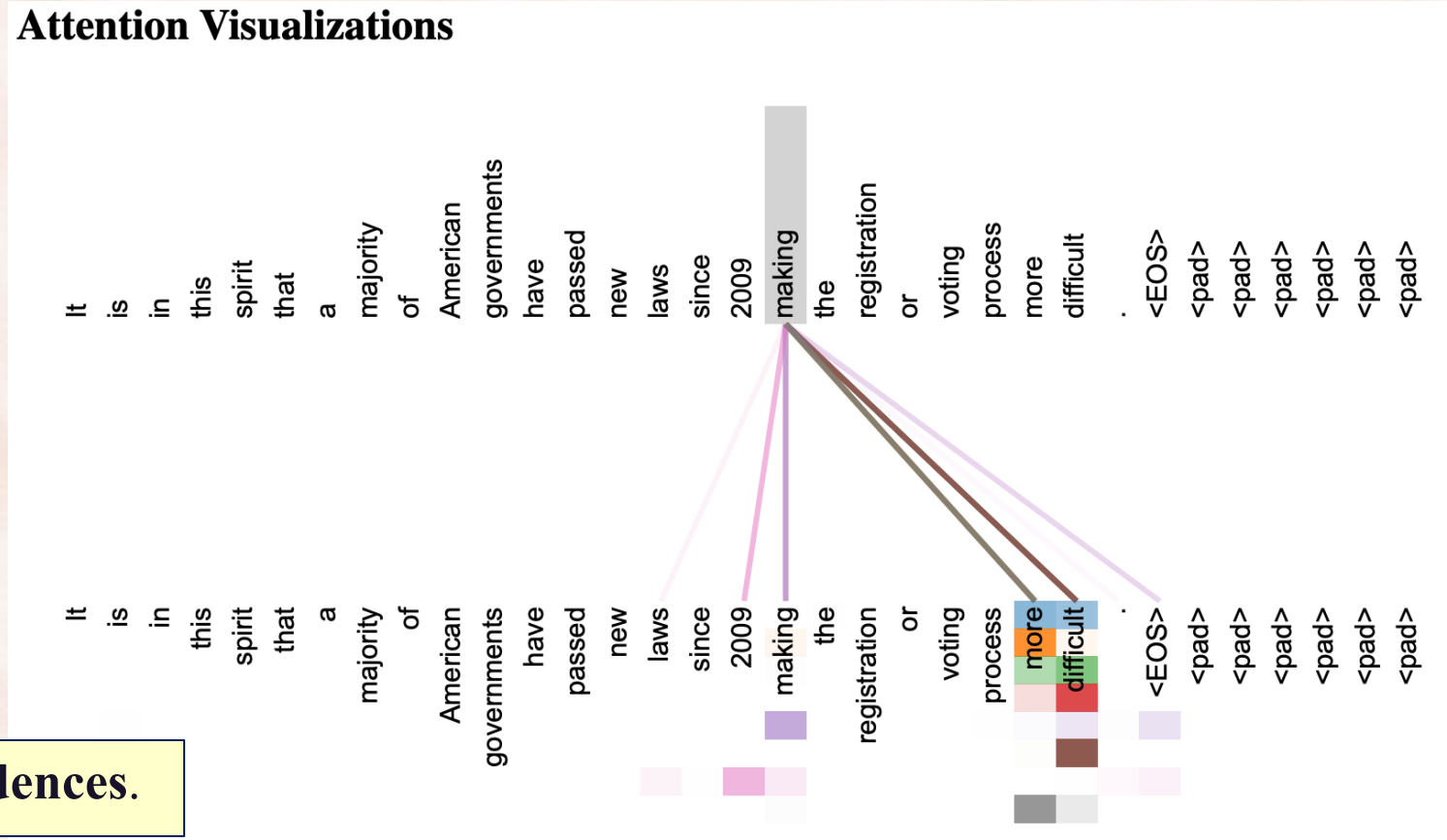
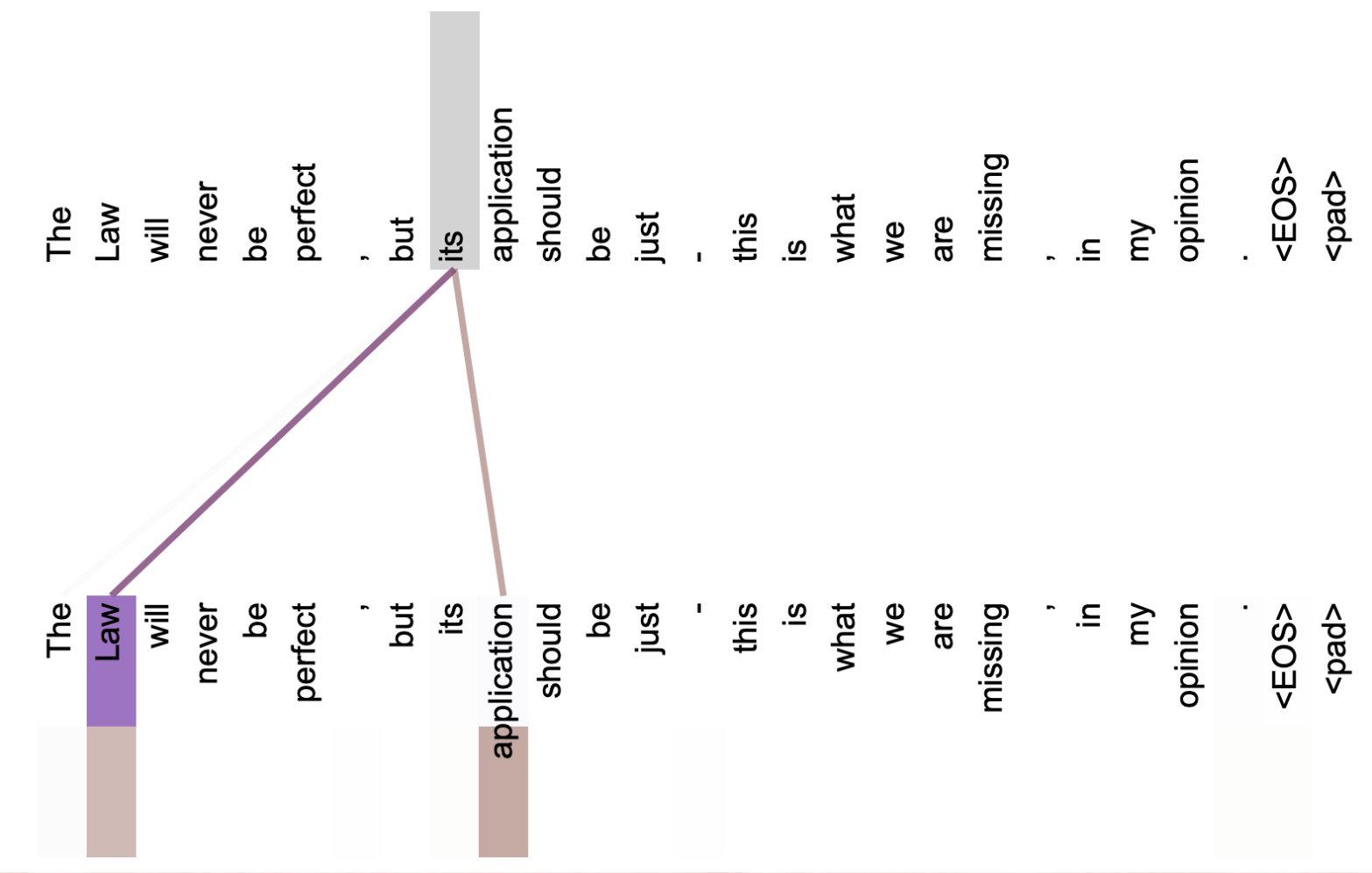


Figure 3: An example of the attention mechanism following long-distance dependencies in the encoder self-attention in layer 5 of 6. Many of the attention heads attend to a distant dependency of the verb 'making', completing the phrase 'making...more difficult'. Attentions here shown only for the word 'making'. Different colors represent different heads. Best viewed in color.

End-to-End: Machine Translation With Transformer



Implicit **coreference resolution**.

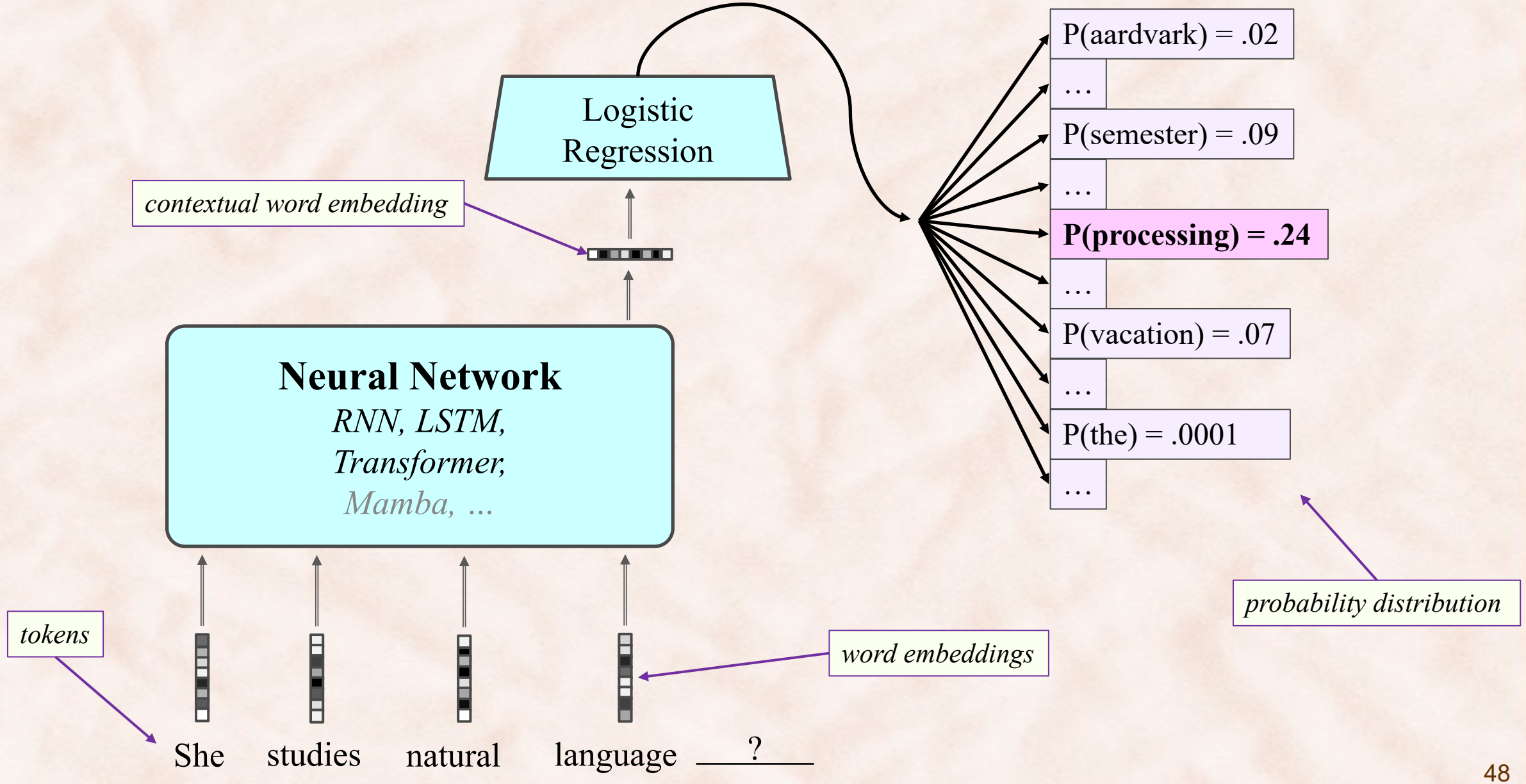
End-to-End: Conditional Text Generation

https://openai.com/blog/better-language-models

- System prompt (human written):
  - *In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.*

- Model completion (machine written):
  - The scientist named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown to science. Now, after almost two centuries, the mystery of what sparked this odd phenomenon is finally solved. Dr. Jorge Pérez, an evolutionary biologist from the University of La Paz, and several companions, were exploring the Andes Mountains when they found a small valley, with no other animals or humans. Pérez noticed that the valley had what appeared to be a natural fountain, surrounded by two peaks of rock and silver snow. Pérez and the others then ventured further into the valley. "By the time we reached the top of one peak, the water looked blue, with some crystals on top," said Pérez …

[*February 14, 2019, introducing GPT-2*]


Logistic Regression
contextual word embedding
Neural Network
RNN, LSTM,
Transformer,
Mamba, …
P(aardvark) = .02
…
P(semester) = .09
…
P(processing) = .24
…
P(vacation) = .07
…
P(the) = .0001
…
probability distribution
tokens
word embeddings
She
studies
natural
language
?

What is Natural Language Processing?

- **Natural Language Processing (NLP)** = developing computer systems that can *process*, *understand*, or *communicate* in natural language (*text* or *speech*):
  - **Natural Languages**: English, Turkish, Japanese, Latin, Hawaiian Creole, Esperanto, American Sign Language, …
    - Music?
  - **Formal Languages**: C++, Java, Python, XML, OWL, Predicate Calculus, Lambda Calculus, …
  - **Natural Languages are significantly more difficult to process than Artificial Languages!**

- What about **Computational Linguistics (CL)**?
  - Computational Linguistics is focused on the *study of language*, using computational tools.
  - NLP is focused on solving language tasks such as *machine translation*, *information extraction*, ***question answering***, *taking instructions*, *holding conversations*, …

CL studies language using NLP/computational tools

[Hamilton et al., EMNLP'16]

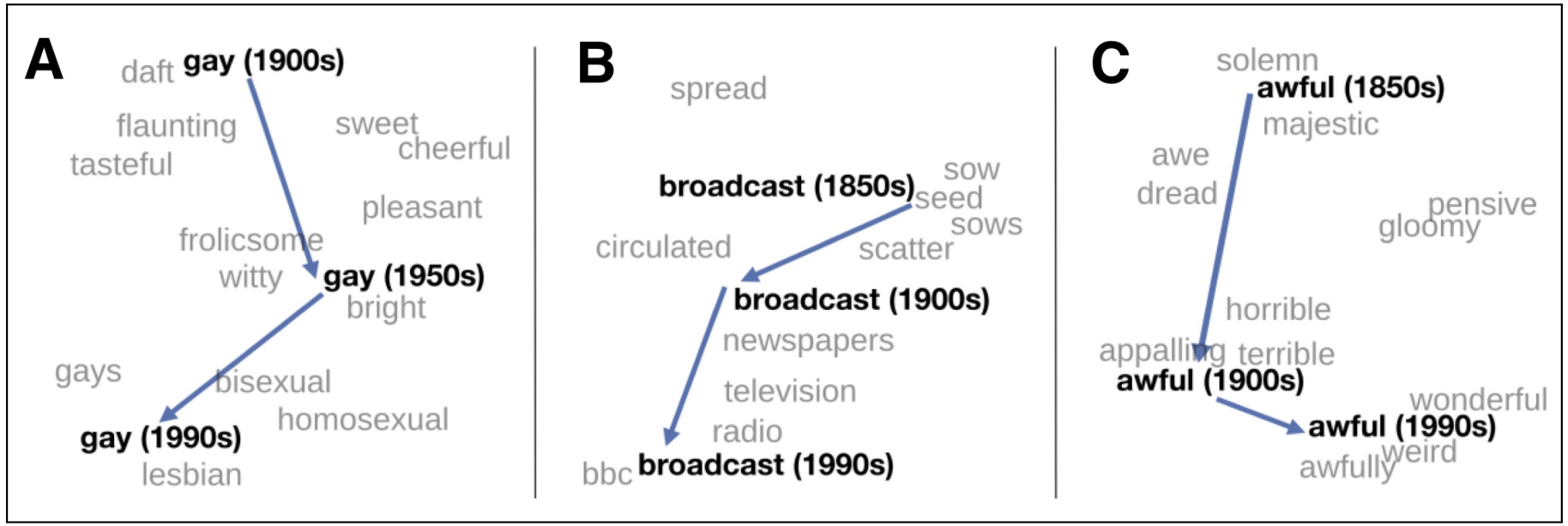


Figure 6.15 A t-SNE visualization of the semantic change of 3 words in English using word2vec vectors. The modern sense of each word, and the grey context words, are computed from the most recent (modern) time-point embedding space. Earlier points are computed from earlier historical embedding spaces. The visualizations show the changes in the word *gay* from meanings related to "cheerful" or "frolicsome" to referring to homosexuality, the development of the modern "transmission" sense of *broadcast* from its original sense of sowing seeds, and the pejoration of the word *awful* as it shifted from meaning "full of awe" to meaning "terrible or appalling" (Hamilton et al., 2016b).

CL studies language using NLP/computational tools

[Hamilton et al., EMNLP'16]

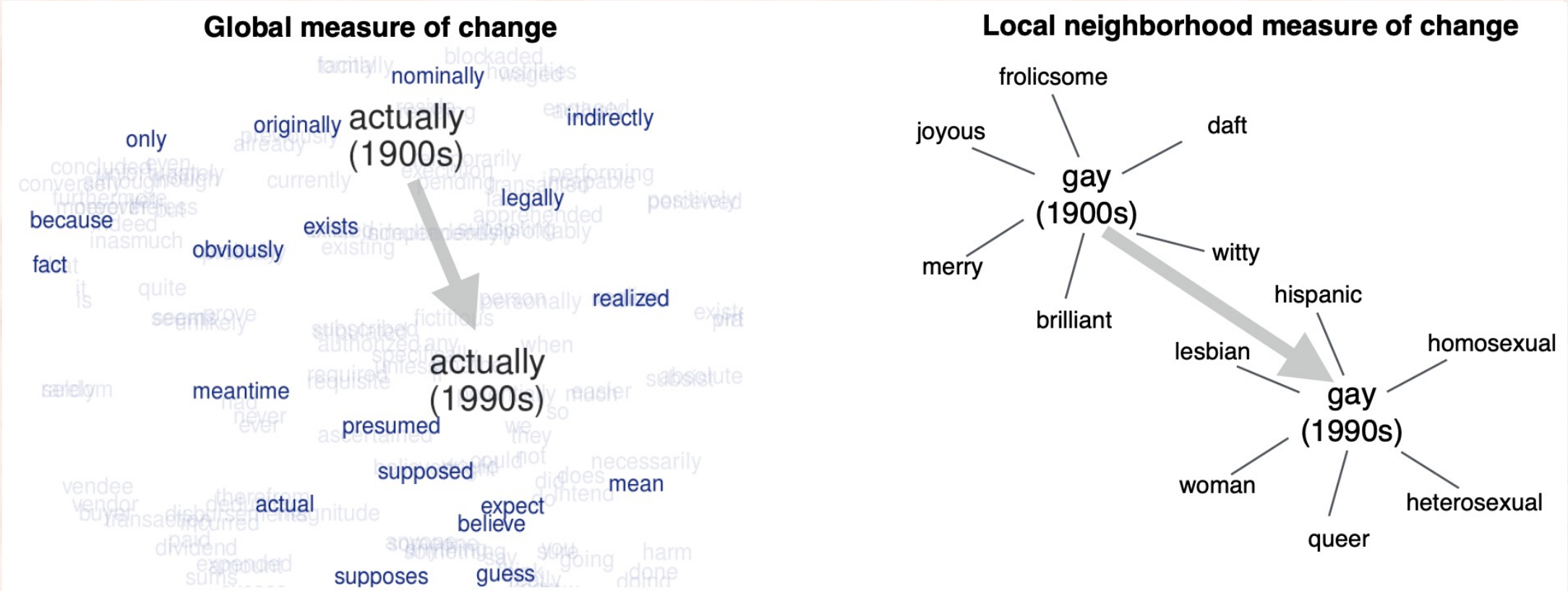


Figure 1: Two different measures of semantic change. With the global measure of change, we measure how far a word has moved in semantic space between two time-periods. This measure is sensitive to subtle shifts in usage and also global effects due to the entire semantic space shifting. For example, this captures how *actually* underwent subjectification during the 20th century, shifting from uses in objective statements about the world ("actually did try") to subjective statements of attitude ("I actually agree"; see Traugott and Dasher, 2001 for details). In contrast, with the local neighborhood measure of change, we measure changes in a word's nearest neighbors, which captures drastic shifts in core meaning, such as *gay*'s shift in meaning over the 20th century.

CL and Psycholinguistics: Syntactic Reduction

- Which of the following utterances are more likely to be produced by speakers?
  1. She created a mobile app dancers like. [example from Meilin Zhan's PhD thesis]
  2. She created a mobile app that dancers like.

  3. My boss thinks I am absolutely crazy. [example from Jaeger T. F.]
  4. My boss thinks that I am absolutely crazy.

  5. My boss confirmed we were absolutely crazy. [example from Jaeger T. F.]
  6. My boss confirmed that we were absolutely crazy.

CL and Psycholinguistics: Syntactic Reduction

- Which of the following utterances are more likely to be produced by speakers?

1. She created a mobile app **dancers** like.
2. 👍 She created a mobile app that dancers like.

red: less predictable (high information).

3. 👍 My boss thinks **I** am absolutely crazy.
4. My boss thinks that I am absolutely crazy.

blue: more predictable (low information).

5. My boss confirmed **we** were absolutely crazy.
6. 👍 My boss confirmed that we were absolutely crazy.

red: less predictable (high information).

- Explained by the **Uniform Information Density (UID)** Hypothesis:

  Speakers aim to convey information at a relatively constant rate (avoid peaks and troughs)

Jaeger, T. F. (2010). Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology*

Required Readings

- Introduction material on parts of speech and syntactic structures:
  - **POS tagging** from Chapter 17, up to and including section 17.2.
  - **Constituency parsing** from Chapter 18, up to and including section 18.3.
  - **Dependency parsing** from Chapter 19, up to and including section 19.1.

- Python introductory lecture slides and language tutorial

- Regular expressions in Python:
  - https://docs.python.org/3/howto/regex.html

- Extracting linguistic features with spaCy:
  - https://spacy.io/usage/linguistic-features