Overview:
Levels of Analysis in NLP

Natural Language Processing

Levels of Language Learning

- When we are growing up, we learn language at multiple levels:
  - Words: what's a real word, how it is spelled
  - Grammar: how words combine to make sentences
  - Meanings: what words and sentences mean in context
  - Inferences: how people draw more out of what is said than what is literally stated

Levels of NLP Analysis

- Similarly, we have different levels in NLP.

Words		Lexical analysis
Grammar		Syntactic analysis
Meanings		Semantic analysis
Inferences		Discourse/entailment analysis

Increasing Difficulty

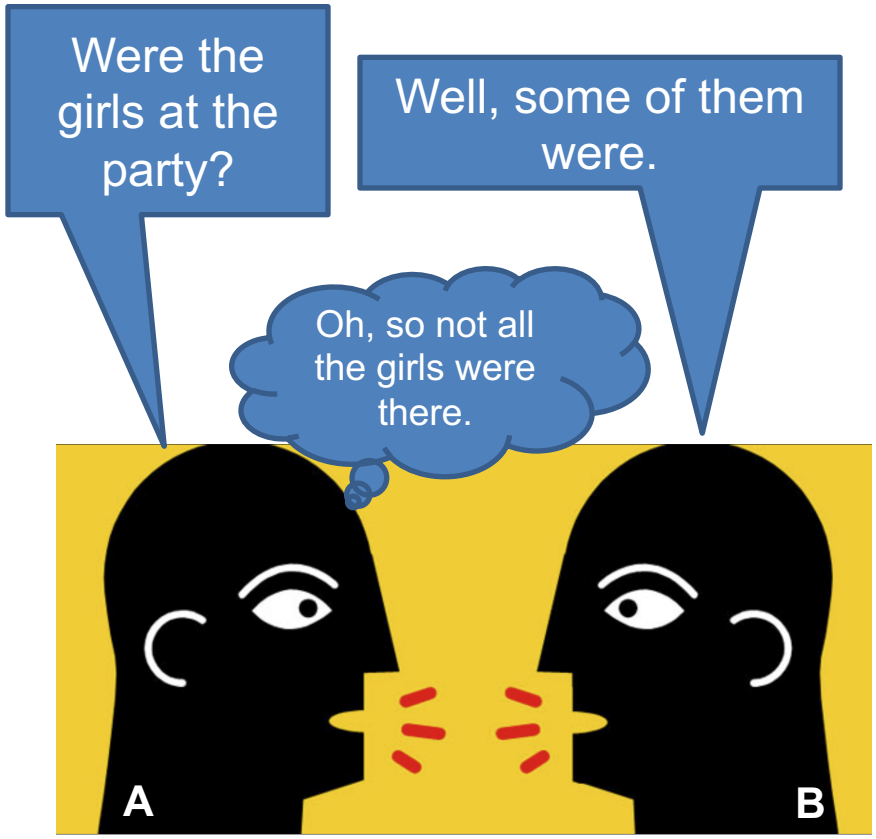
As human beings, we find it gets harder to learn a language as you progress to the next level.

Suppose you are an English speaker learning German for the first time.

Word: “Attention!”	⟷	“Achtung!”
Grammar: “It’s raining.”	⟷	“Es regnet.” (lit: “It rains.”)
Meanings: “What time is it? Quarter ’til nine!”	⟷	“Wieviel Uhr ist es? Dreiviertel zehn!” (lit: How much clock is it? Three-quarters ten!)

Analysis at the Level of Discourse

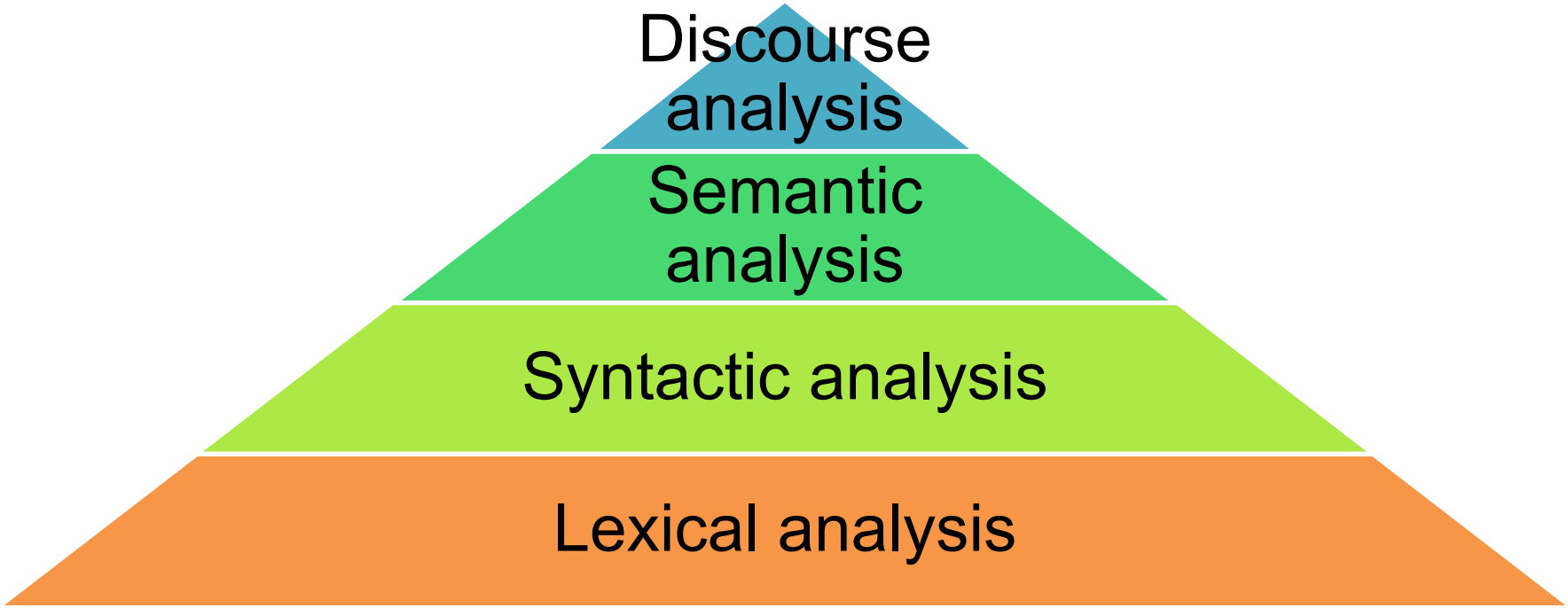
The inferences we make in discourse are the hardest of all. Consider the following:



Speaker A makes an inference. Is it a valid inference?

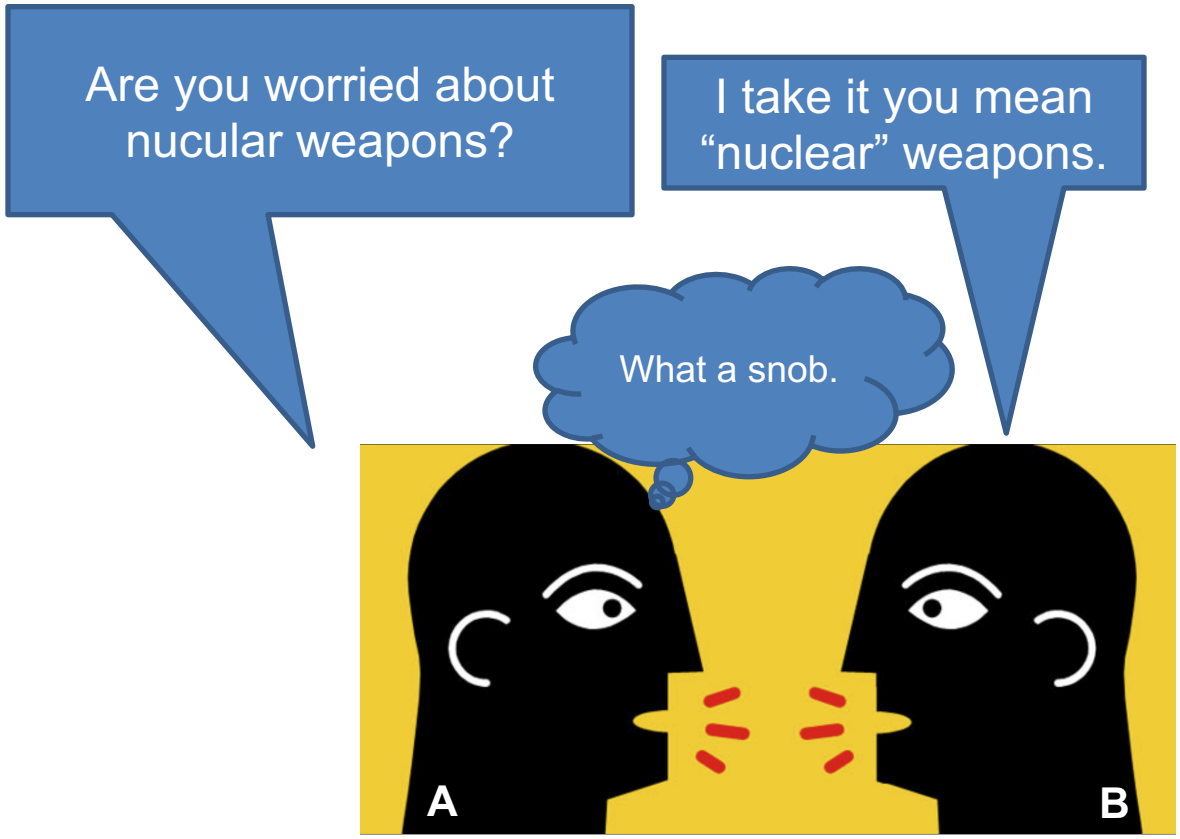
Levels of Analysis

It is tempting to think of these levels of analysis as layers, such that each one builds upon the lower levels beneath it.



Multilevel Analysis

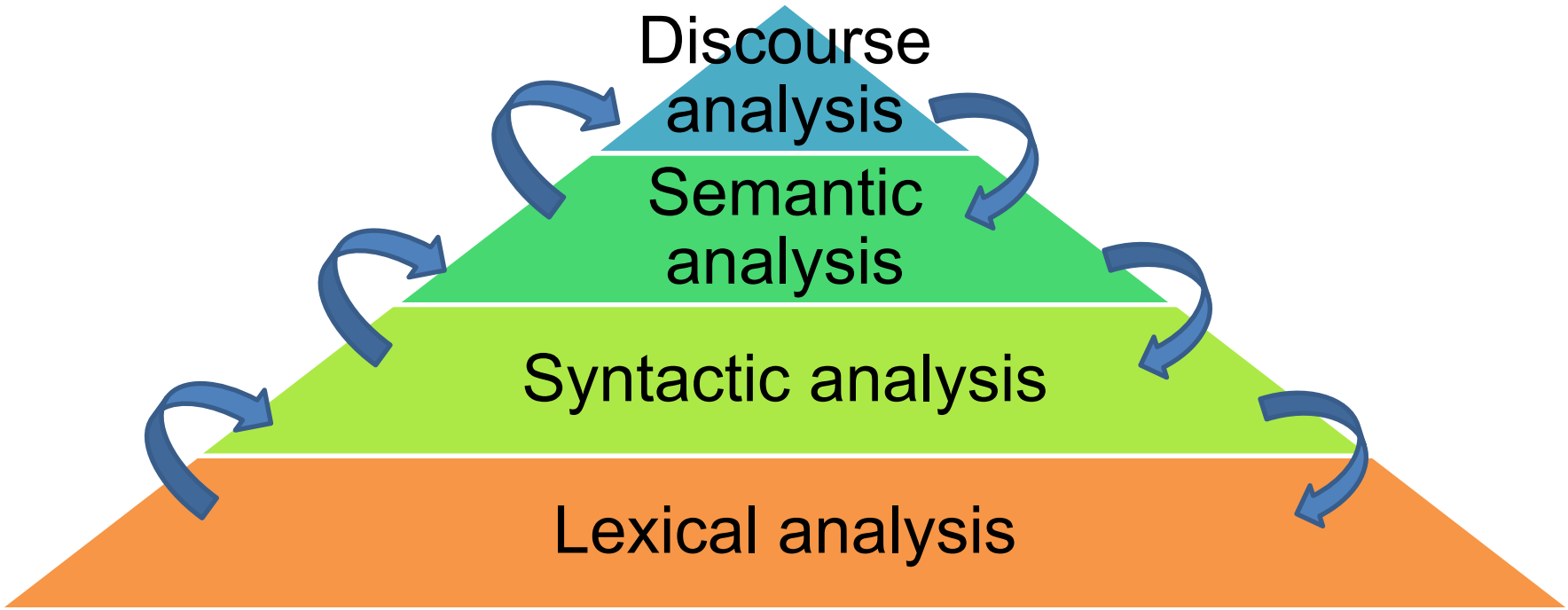
However, in real life, we use all the levels to help us understand all the levels:



What level of analysis was speaker B correcting? What level of analysis was he using to make that correction?

Levels of Analysis

So, for convenience, we can continue to think of these as higher and lower levels as long as we remember that they have every manner of interconnection.



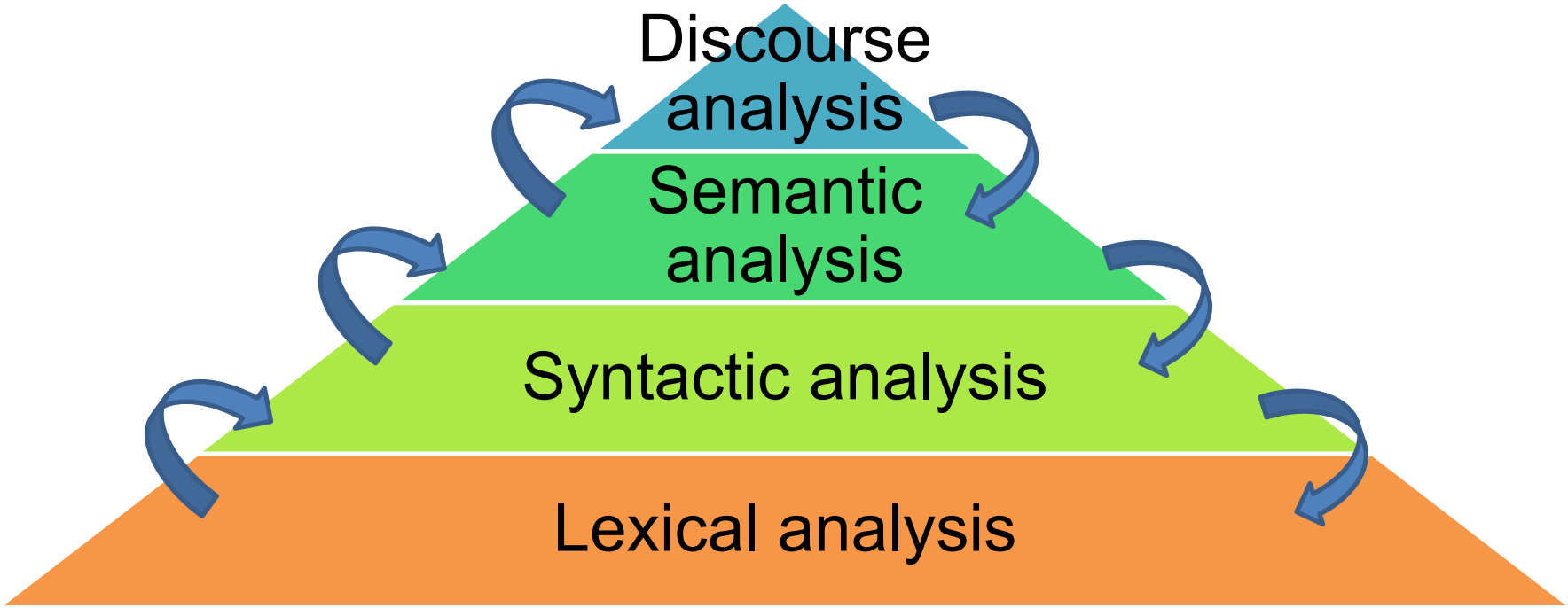
DataScience@SMU

Overview: Lexical Analysis in NLP

Natural Language Processing

Levels of Analysis

For convenience, we can think of there being higher and lower levels of analysis in NLP as long as we remember they are interconnected.



Morphology and Stemming

Morphology means the study of “morphemes,” which are the units that our words are made of.

- Take the word “running.” The root word “run” is a morpheme, and the often-used “-ing” ending is itself a morpheme.

Making an algorithm that when inputted “running” automatically gives you the root morpheme “run” is called a “stemmer.”

What applications could benefit from a stemmer?

Attachment of Metadata

Corpus-derived metadata of various types can make valuable enhancements to lexical resources. For each headword consider adding:

- Word-frequency score
- Common collocations
- Common co-occurring words, context words, etc.

A word-cloud built from co-occurrence frequencies with the word “MORALS” might look like this.

Collocations

Collocations can make a big impact on the meaning of the main word. Consider all these collocations of “take.”

take a shower	take charge	take a seat
take care	take responsibility	take heart
take a class	take a look	take an opportunity
take damage	take an exam	take notes
take a role	take a nap	take notice
take initiative	take a claim	take a picture
take a break	take a dump	take a taxi
take the lead	take a rest	take action
take a chance	take a class	take advantage

It may be better to have a headword entry in the lexicon for each of these—then make sure that lexical look-ups are done right!

Enumeration of Word Senses

- Most words are "polysemous," meaning they can have more than one meaning.
- Usually, a lexicon tries to list word senses in order, from most common to least common in occurrences.

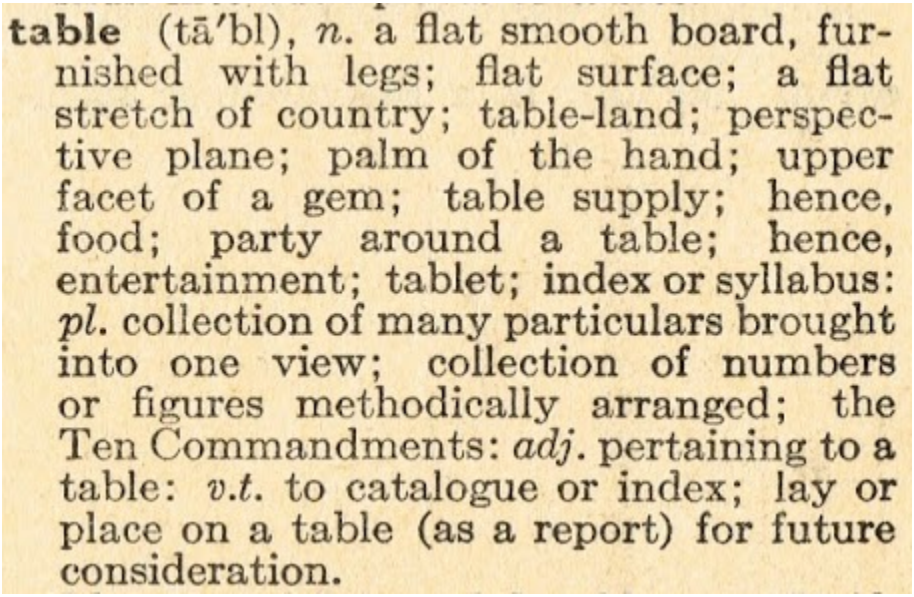


table (tā′bl), *n.* a flat smooth board, furnished with legs; flat surface; a flat stretch of country; table-land; perspective plane; palm of the hand; upper facet of a gem; table supply; hence, food; party around a table; hence, entertainment; tablet; index or syllabus: *pl.* collection of many particulars brought into one view; collection of numbers or figures methodically arranged; the Ten Commandments: *adj.* pertaining to a table: *v.t.* to catalogue or index; lay or place on a table (as a report) for future consideration.

Enumeration of Word Senses

If WordNet is “the Bible” here, then many real-world word senses are “apocryphal.”

WordNet Search - 3.1
- WordNet home page - Glossary - Help

Word to search for: run Search WordNet

Display Options: (Select option to change) Change

Key: "S:" = Show Synset (semantic) relations, "W:" = Show Word (lexical) relations

Display options for sense: (gloss) "an example sentence"

Noun

- S: (n) **run**, tally (a score in baseball made by a runner touching all four bases safely) *"the Yankees scored 3 runs in the bottom of the 9th"; "their first tally came in the 3rd inning"*
- S: (n) test, trial, **run** (the act of testing something) *"in the experimental trials the amount of carbon was measured separately"; "he called each flip of the coin a new trial"*
- S: (n) footrace, foot race, **run** (a race run on foot) *"she broke the record for the half-mile run"*
- S: (n) streak, **run** (an unbroken series of events) *"had a streak of bad luck"; "Nicklaus had a run of birdies"*
- S: (n) **run**, running, running play, running game ((American football) a play in which a player attempts to carry the ball through or past the opposing team) *"the defensive line braced to stop the run"; "the coach put great emphasis on running"*
- S: (n) **run** (a regular trip) *"the ship made its run in record time"*
- S: (n) **run**, running (the act of running; traveling on foot at a fast pace) *"he broke into a run"; "his daily run keeps him fit"*
- S: (n) **run** (the continuous period of time during which something (a machine or a factory) operates or continues in operation) *"the assembly line was on a 12-hour run"*
- S: (n) **run** (unrestricted freedom to use) *"he has the run of the house"*
- S: (n) **run** (the production achieved during a continuous period of operation (of a machine or factory etc.)) *"a daily run of 100,000 gallons of paint"*

Domain Relations

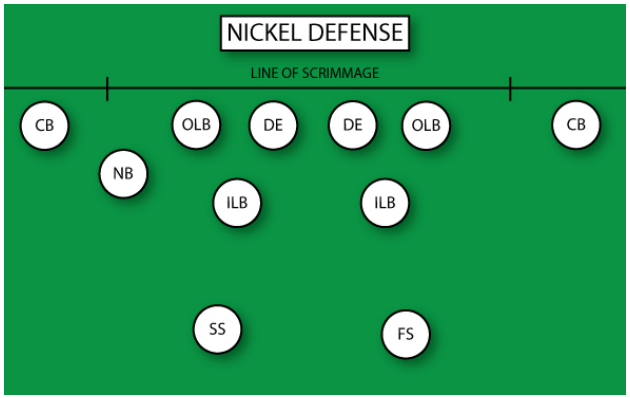
- Once we have word senses, we realize some senses are linked to, or much stronger in, one domain vs. another.
- Consider “nickel” as a metal, a coin, a sports term.

- - - - - - - D O M A I N S - - - - - - -

MONEY

METALLURGY

SPORTS



Text and Corpus Analytics

At the lexical level, we can already do a lot of very interesting analysis of large texts, and even of whole corpora.
Examples:

- Spell correction
- Terminology extraction
- Lexical diversity measurement

```
>>> lexical_diversity(text3)
0.06230453042623537
>>> lexical_diversity(text5)
0.13477005109975562
```

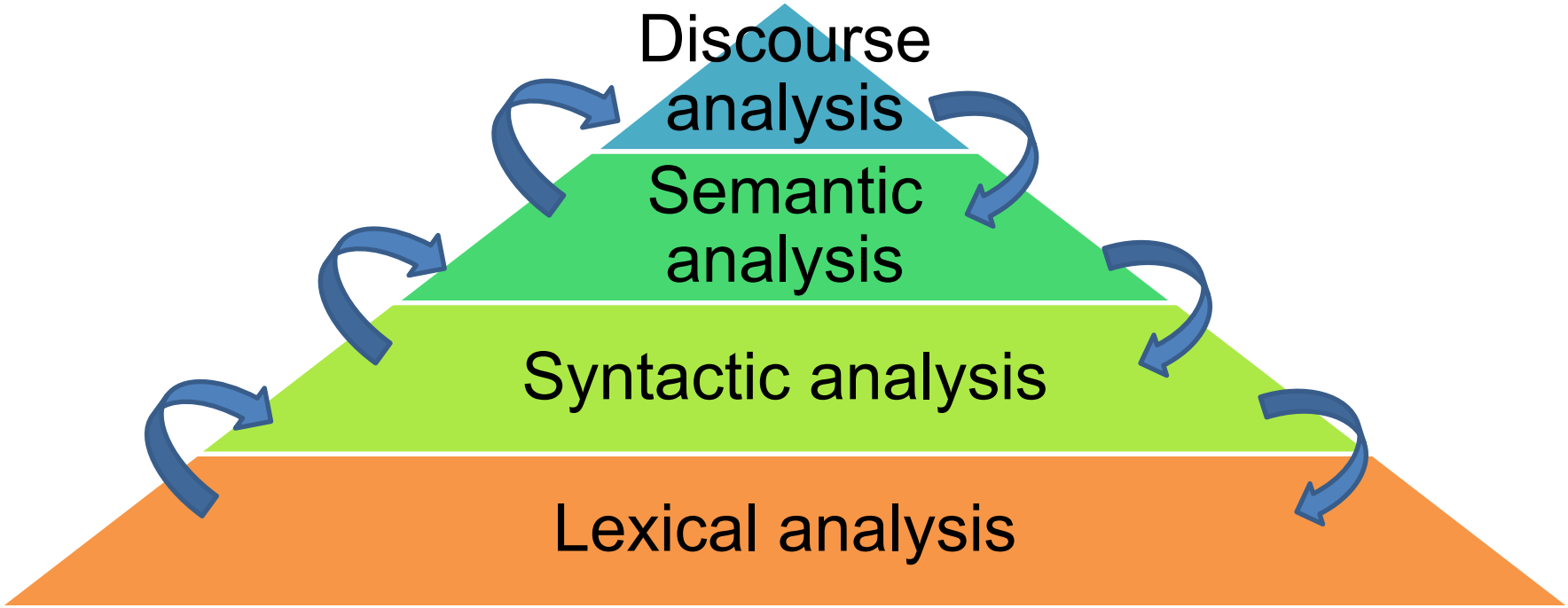
DataScience@SMU

Overview: Lexical Analysis in NLP

Natural Language Processing

Levels of Analysis

For convenience, we can think of there being higher and lower levels of analysis in NLP as long as we remember they are interconnected.



Morphology and Stemming

Morphology means the study of “morphemes,” which are the units that our words are made of.

- Take the word “running.” The root word “run” is a morpheme, and the often-used “-ing” ending is itself a morpheme.

Making an algorithm that when inputted “running” automatically gives you the root morpheme “run” is called a “stemmer.”

What applications could benefit from a stemmer?

Attachment of Metadata

Corpus-derived metadata of various types can make valuable enhancements to lexical resources. For each headword consider adding:

- Word-frequency score
- Common collocations
- Common co-occurring words, context words, etc.

A word-cloud built from co-occurrence frequencies with the word “MORALS” might look like this.

Collocations

Collocations can make a big impact on the meaning of the main word. Consider all these collocations of “take.”

take a shower	take charge	take a seat
take care	take responsibility	take heart
take a class	take a look	take an opportunity
take damage	take an exam	take notes
take a role	take a nap	take notice
take initiative	take a claim	take a picture
take a break	take a dump	take a taxi
take the lead	take a rest	take action
take a chance	take a class	take advantage

It may be better to have a headword entry in the lexicon for each of these—then make sure that lexical look-ups are done right!

Enumeration of Word Senses

- Most words are "polysemous," meaning they can have more than one meaning.
- Usually, a lexicon tries to list word senses in order, from most common to least common in occurrences.

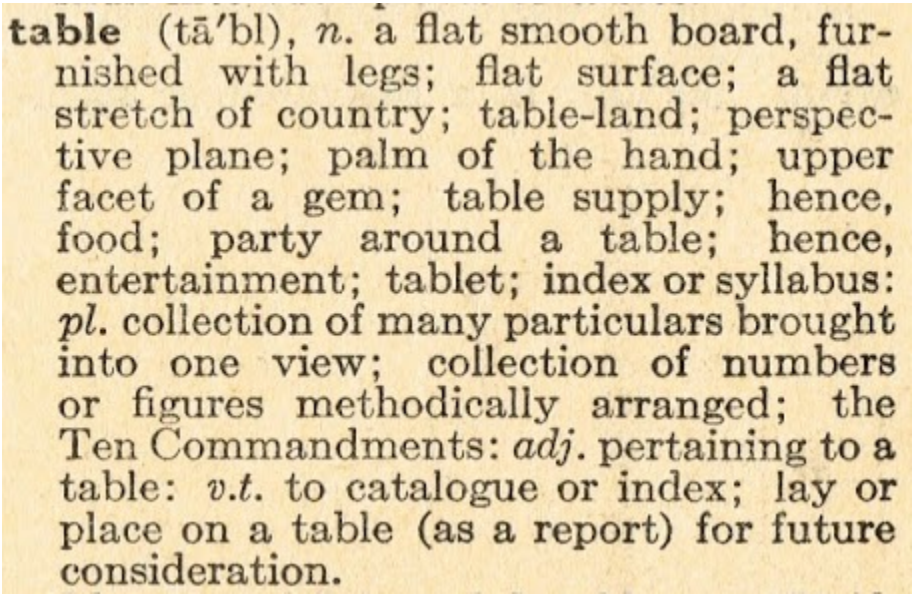


table (tā′bl), *n.* a flat smooth board, furnished with legs; flat surface; a flat stretch of country; table-land; perspective plane; palm of the hand; upper facet of a gem; table supply; hence, food; party around a table; hence, entertainment; tablet; index or syllabus: *pl.* collection of many particulars brought into one view; collection of numbers or figures methodically arranged; the Ten Commandments: *adj.* pertaining to a table: *v.t.* to catalogue or index; lay or place on a table (as a report) for future consideration.

Enumeration of Word Senses

If WordNet is “the Bible” here, then many real-world word senses are “apocryphal.”

WordNet Search - 3.1
- WordNet home page - Glossary - Help

Word to search for: run Search WordNet

Display Options: (Select option to change) Change

Key: "S:" = Show Synset (semantic) relations, "W:" = Show Word (lexical) relations

Display options for sense: (gloss) "an example sentence"

Noun

- S: (n) **run**, tally (a score in baseball made by a runner touching all four bases safely) *"the Yankees scored 3 runs in the bottom of the 9th"; "their first tally came in the 3rd inning"*
- S: (n) test, trial, **run** (the act of testing something) *"in the experimental trials the amount of carbon was measured separately"; "he called each flip of the coin a new trial"*
- S: (n) footrace, foot race, **run** (a race run on foot) *"she broke the record for the half-mile run"*
- S: (n) streak, **run** (an unbroken series of events) *"had a streak of bad luck"; "Nicklaus had a run of birdies"*
- S: (n) **run**, running, running play, running game ((American football) a play in which a player attempts to carry the ball through or past the opposing team) *"the defensive line braced to stop the run"; "the coach put great emphasis on running"*
- S: (n) **run** (a regular trip) *"the ship made its run in record time"*
- S: (n) **run**, running (the act of running; traveling on foot at a fast pace) *"he broke into a run"; "his daily run keeps him fit"*
- S: (n) **run** (the continuous period of time during which something (a machine or a factory) operates or continues in operation) *"the assembly line was on a 12-hour run"*
- S: (n) **run** (unrestricted freedom to use) *"he has the run of the house"*
- S: (n) **run** (the production achieved during a continuous period of operation (of a machine or factory etc.)) *"a daily run of 100,000 gallons of paint"*

Domain Relations

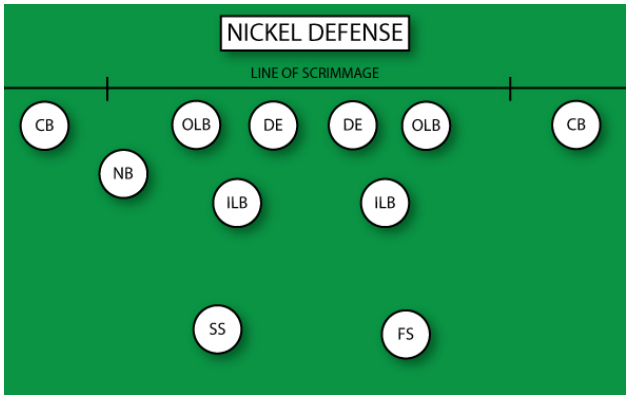
- Once we have word senses, we realize some senses are linked to, or much stronger in, one domain vs. another.
- Consider “nickel” as a metal, a coin, a sports term.

- - - - - - - D O M A I N S - - - - - - -

MONEY

METALLURGY

SPORTS



Text and Corpus Analytics

At the lexical level, we can already do a lot of very interesting analysis of large texts, and even of whole corpora.
Examples:

- Spell correction
- Terminology extraction
- Lexical diversity measurement

```
>>> lexical_diversity(text3)
0.06230453042623537
>>> lexical_diversity(text5)
0.13477005109975562
```

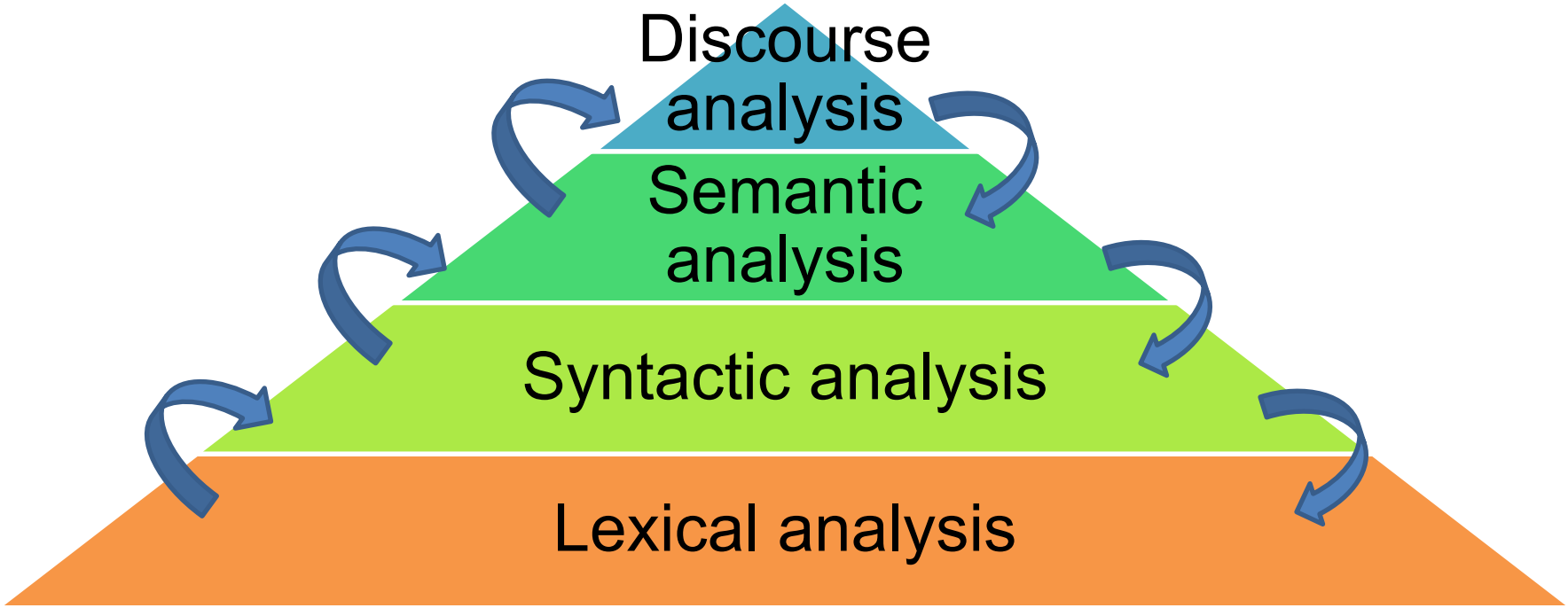
DataScience@SMU

Overview: Syntactic Analysis in NLP

Natural Language Processing

Levels of Analysis

For convenience, we can think of there being higher and lower levels of analysis in NLP as long as we remember they are interconnected.

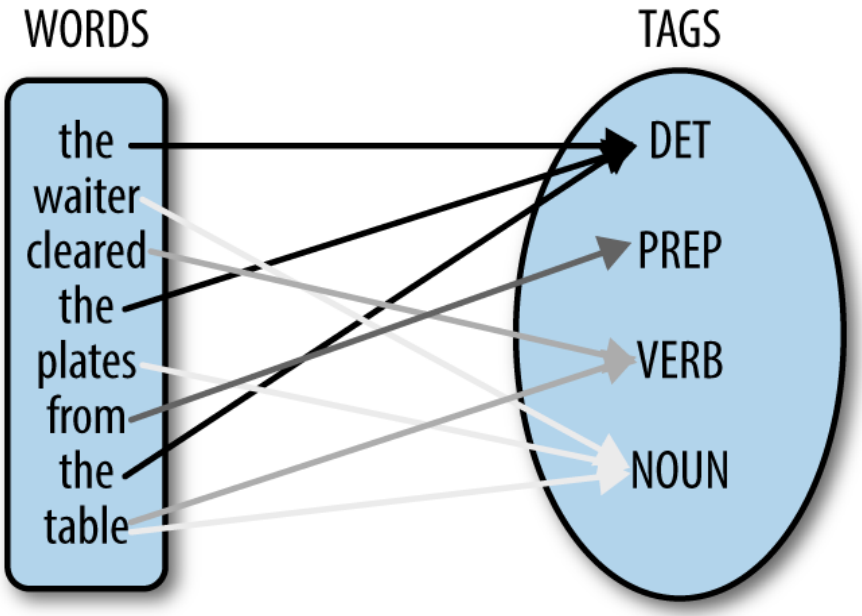


Sentence Boundary Detection

- The most common sentence boundary marker, the period, “.”, is used unfortunately for abbreviations; therefore, sentence boundary detection is nontrivial in English.

POS Tagging

- Most words can have more than one part of speech; therefore, POS tagging is nontrivial.



POS Tagging

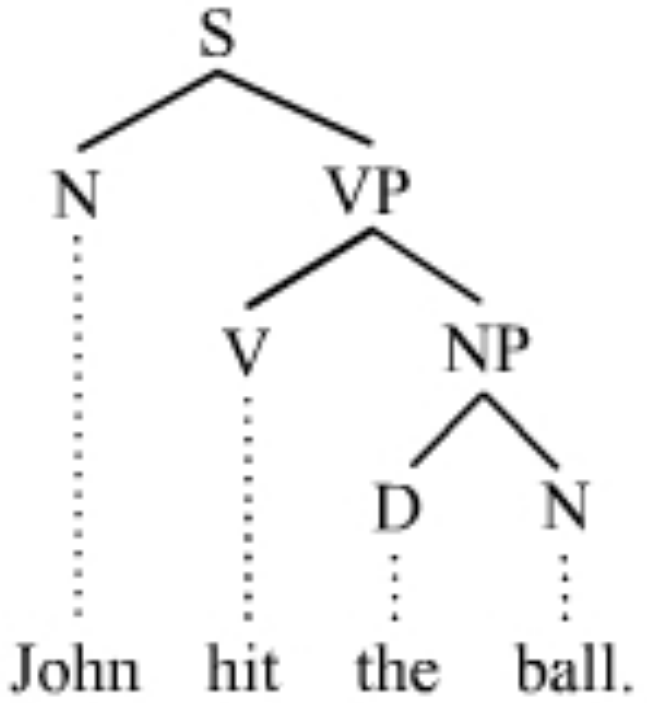
One of the most commonly used tag sets:

Penn Treebank Tagset

Tag	Description	Example	Tag	Description	Example
CC	Coordin. Conjunction	*and, but, or*	SYM	Symbol	+,%, &
CD	Cardinal number	*one, two, three*	TO	"to"	*to*
DT	Determiner	*a, the*	UH	Interjection	*ah, oops*
EX	Existential 'there'	*there*	VB	Verb, base form	*eat*
FW	Foreign word	*mea culpa*	VBD	Verb, past tense	*ate*
IN	Preposition/sub-conj	*of, in, by*	VBG	Verb, gerund	*eating*
JJ	Adjective	*yellow*	VBN	Verb, past participle	*eaten*
JJR	Adj., comparative	*bigger*	VBP	Verb, non-3sg pres	*eat*
JJS	Adj., superlative	*wildest*	VBZ	Verb, 3sg pres	*eats*
LS	List item marker	*1, 2, One*	WDT	Wh-determiner	*which, that*
MD	Modal	*can, should*	WP	Wh-pronoun	*what, who*
NN	Noun, sing. or mass	*llama*	WP$	Possessive wh-	*whose*
NNS	Noun, plural	*llamas*	WRB	Wh-adverb	*how, where*
NNP	Proper noun, singular	*IBM*	$	Dollar sign	*$*
NNPS	Proper noun, plural	*Carolinas*	#	Pound sign	#
PDT	Predeterminer	*all, both*	"	Left quote	(' or ")
POS	Possessive ending	*'s*	"	Right quote	(' or ")
PRP	Personal pronoun	*I, you, he*	(	Left parenthesis	([, (, {, <)
PRP$	Possessive pronoun	*your, one's*	)	Right parenthesis	( ],), }, >)
RB	Adverb	*quickly, never*	,	Comma	,
RBR	Adverb, comparative	*faster*	.	Sentence-final punc	(. ! ?)
RBS	Adverb, superlative	*fastest*	:	Mid-sentence punc	(: ; ... – -)
RP	Particle	*up, off*			

Parsing

- No parser is perfect, but for most applications we can live with less-than-perfect parsing.



Lemmatization

- A lemma is the canonical (conventional) form that represents a set of related word forms, e.g., for *run, runs, ran, running,* the lemma is "run."
- Often, a context-free stemming gives us the lemma, but in some cases not. Example:

  `wetter` → `wet`

  but

  `better` → `bet?`

  Usually not!

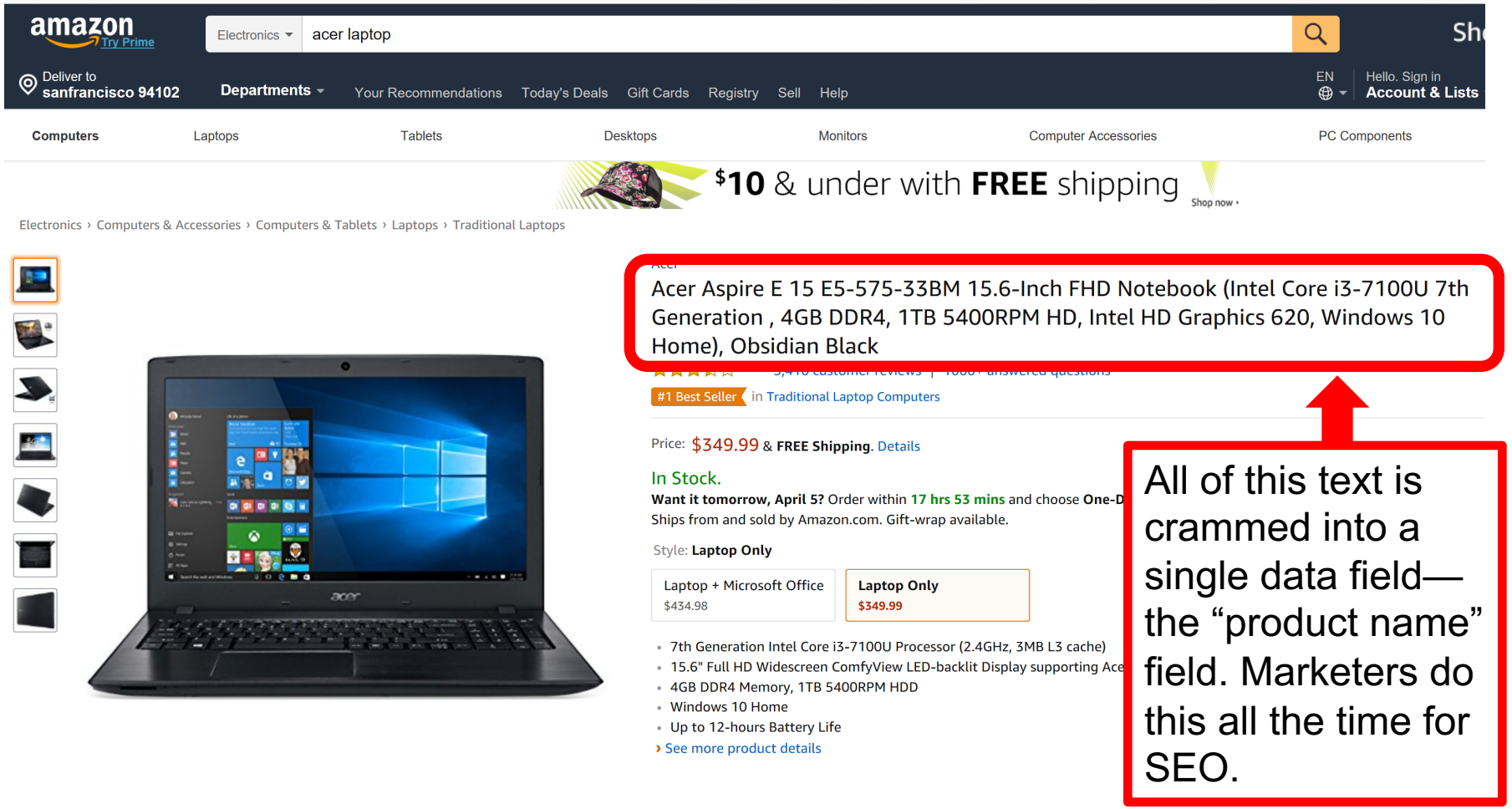
  `better` → `good`

- If we parse "better" as an adjective, not a noun, then we can determine its correct lemma.

Discrete Text Field Analysis

- Unitizing
- Normalizing
- “Smart ETL”

Discrete Text Field Analysis



These text strings obviously do not follow regular grammar, so a standard parser is no help to us. We might write a custom parser using a catalog of product names and features.

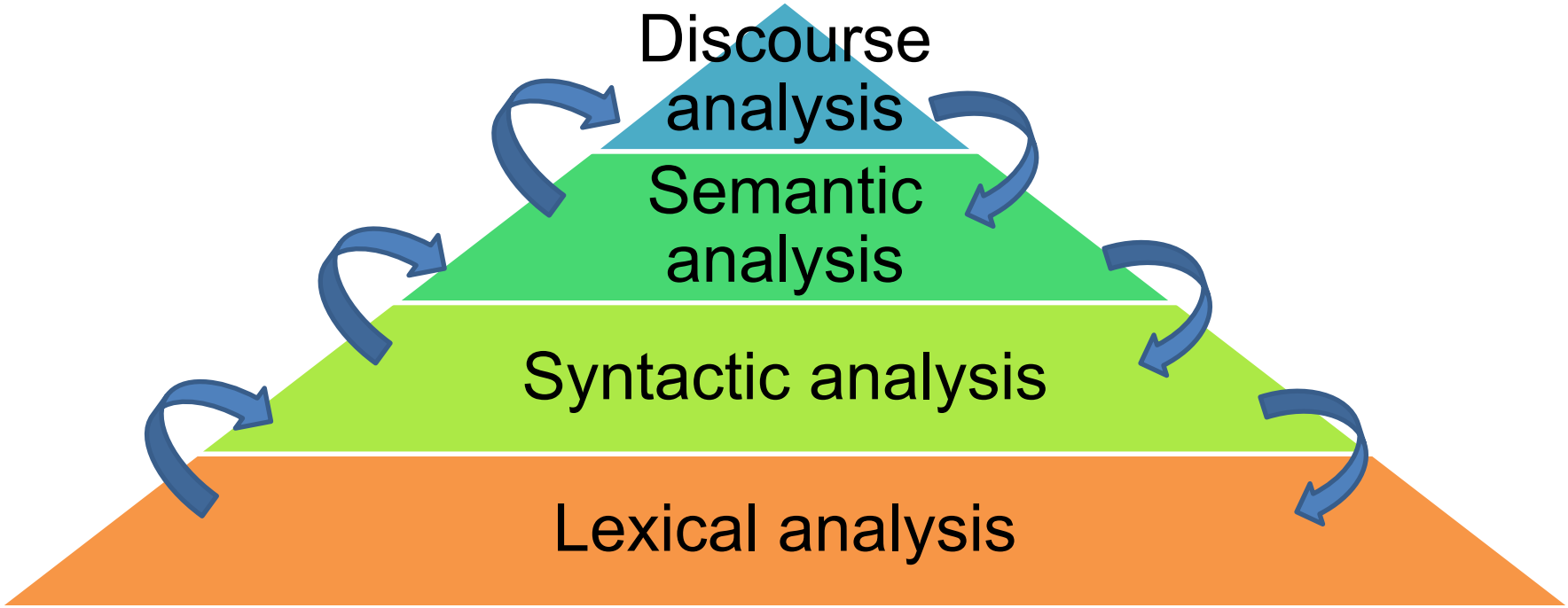
DataScience@SMU

Overview: Semantic Analysis in NLP

Natural Language Processing

Levels of Analysis

For convenience, we can think of there being higher and lower levels of analysis in NLP as long as we remember they are interconnected.



Semantic Analysis

Some major types:

- Named entity extraction (NEE a.k.a. NER)
- Relationship extraction (between NEs)
- Word sense disambiguation (WSD)
- Classification
- Tagging
- Topic segmentation
- Sentiment analysis

Named-Entity Extraction

- Also called “named-entity recognition” (abbreviated NEE or NER).
- Simple NEE just recognizes entities without typing.
- More sophisticated NEE does basic typing:
  - Persons
  - Organizations
  - Places
  - Events

Named-Entity Extraction

Good NEE will cluster together many variants, including epithets.

Hillary Clinton	Donald Trump
Hillary Rodham Clinton	The Donald
Senator Clinton	Donald J. Trump
Secretary Clinton	Donald Jackass Trump
Hitlery Clinton	Donald Drumpf
Crooked Hillary	Herr Trump

Relationship Extraction (between NEs)

- What relation does one NE have to another? Sometimes we can answer this during the NEE process.

These are sometimes called "triples" because there are three elements: two NEs and one relation.

Collections of triples are called "triple stores."

Word Sense Disambiguation (WSD)

WordNet Search - 3.1

- WordNet home page - Glossary - Help

Word to search for: chair Search WordNet

Display Options: (Select option to change) Change

Key: "S:" = Show Synset (semantic) relations, "W:" = Show Word (lexical) relations

Display options for sense: (gloss) "an example sentence"

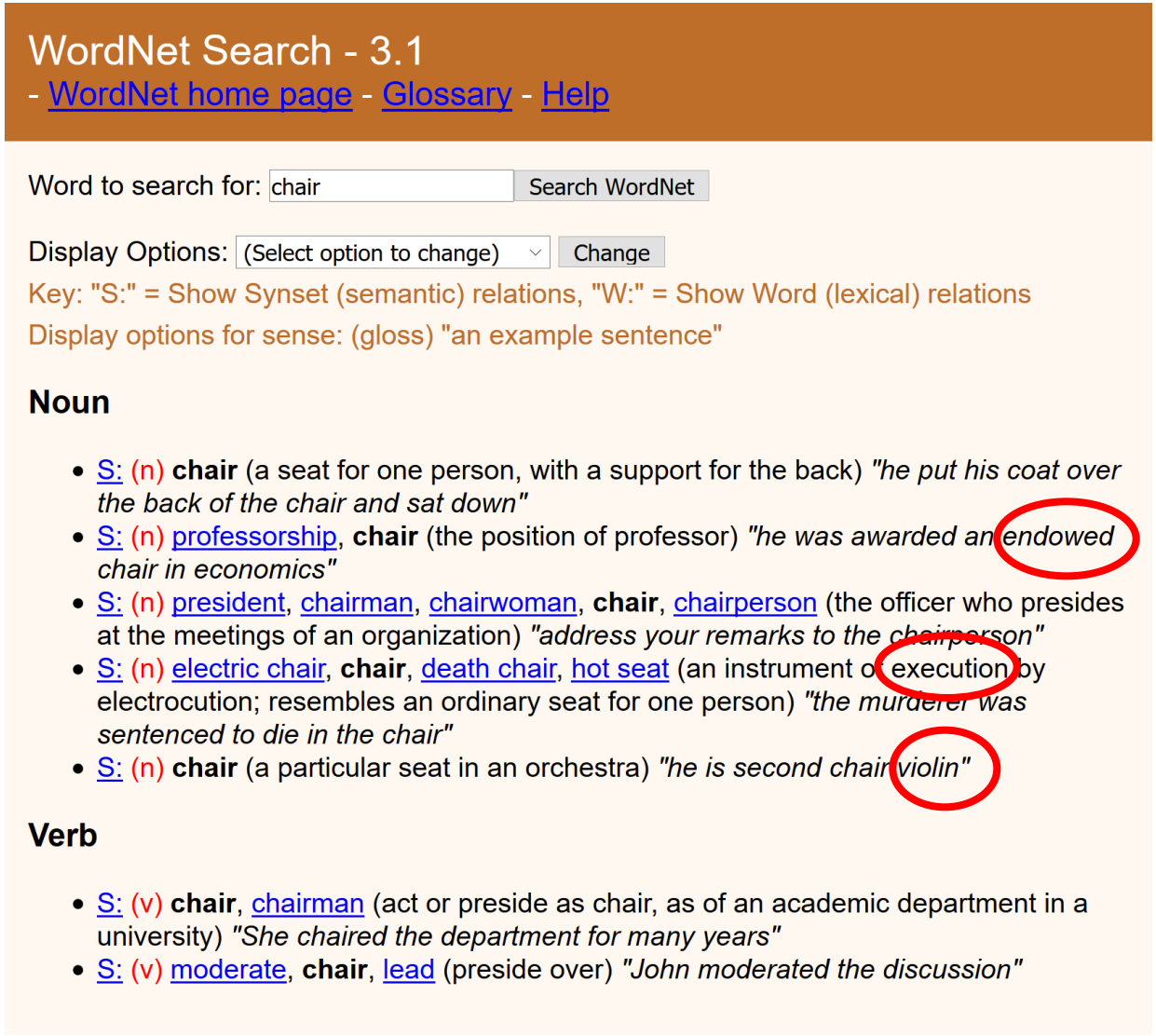
Noun

- S: (n) **chair** (a seat for one person, with a support for the back) *"he put his coat over the back of the chair and sat down"*
- S: (n) professorship, **chair** (the position of professor) *"he was awarded an endowed chair in economics"*
- S: (n) president, chairman, chairwoman, **chair**, chairperson (the officer who presides at the meetings of an organization) *"address your remarks to the chairperson"*
- S: (n) electric chair, **chair**, death chair, hot seat (an instrument of execution by electrocution; resembles an ordinary seat for one person) *"the murderer was sentenced to die in the chair"*
- S: (n) **chair** (a particular seat in an orchestra) *"he is second chair violin"*

Verb

- S: (v) **chair**, chairman (act or preside as chair, as of an academic department in a university) *"She chaired the department for many years"*
- S: (v) moderate, **chair**, lead (preside over) *"John moderated the discussion"*

Word Sense Disambiguation (WSD)



WordNet Search - 3.1
- WordNet home page - Glossary - Help

Word to search for: chair Search WordNet

Display Options: (Select option to change) Change

Key: "S:" = Show Synset (semantic) relations, "W:" = Show Word (lexical) relations

Display options for sense: (gloss) "an example sentence"

Noun

- S: (n) **chair** (a seat for one person, with a support for the back) *"he put his coat over the back of the chair and sat down"*
- S: (n) professorship, **chair** (the position of professor) *"he was awarded an endowed chair in economics"*
- S: (n) president, chairman, chairwoman, **chair**, chairperson (the officer who presides at the meetings of an organization) *"address your remarks to the chairperson"*
- S: (n) electric chair, **chair**, death chair, hot seat (an instrument of execution by electrocution; resembles an ordinary seat for one person) *"the murderer was sentenced to die in the chair"*
- S: (n) **chair** (a particular seat in an orchestra) *"he is second chair violin"*

Verb

- S: (v) **chair**, chairman (act or preside as chair, as of an academic department in a university) *"She chaired the department for many years"*
- S: (v) moderate, **chair**, lead (preside over) *"John moderated the discussion"*

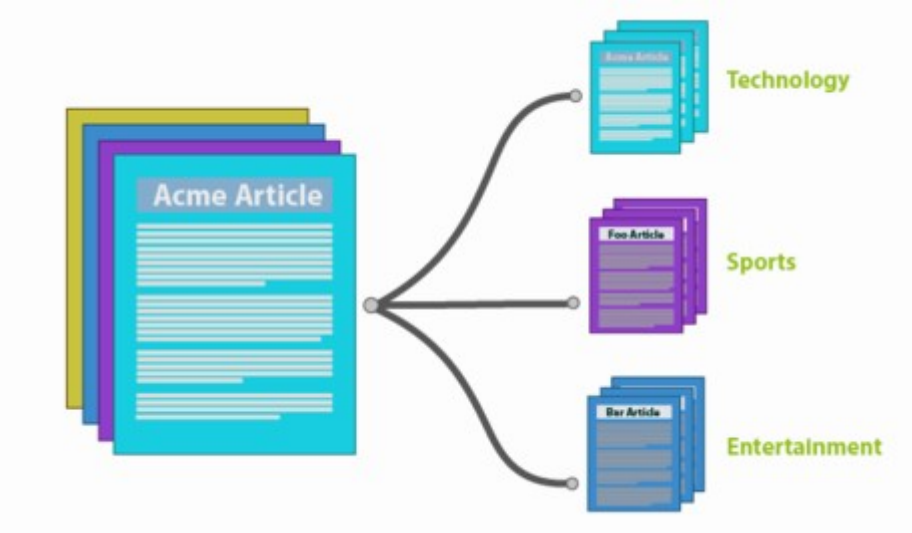
Context words give us a chance to disambiguate intended senses.

Doing this at scale is hard! It's still an open research area of NLP (i.e., unsolved).

Classification

- In classification, we use a tree-structured graph to place documents into categories.
- Often we employ machine learning methods, such as SVM*.

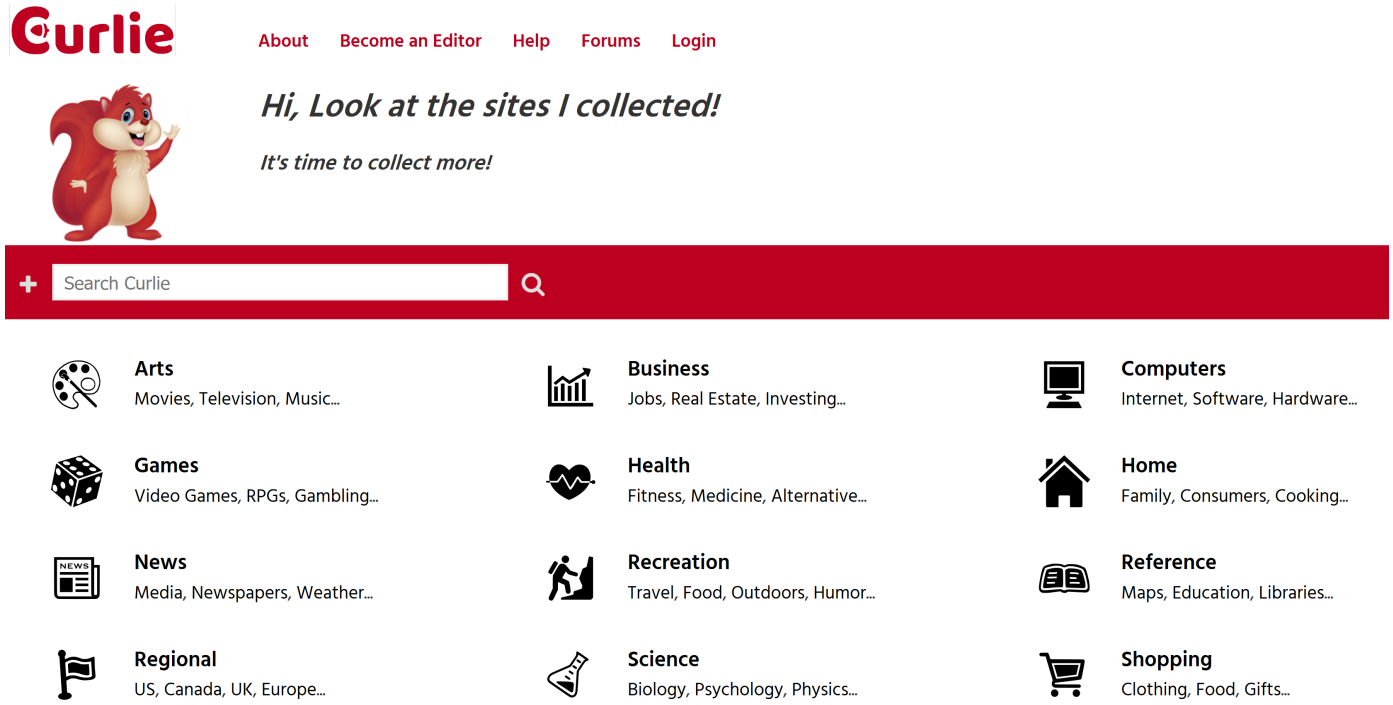
We always start from a preexisting category schema, often called a "taxonomy."



*support vector machine

Classification

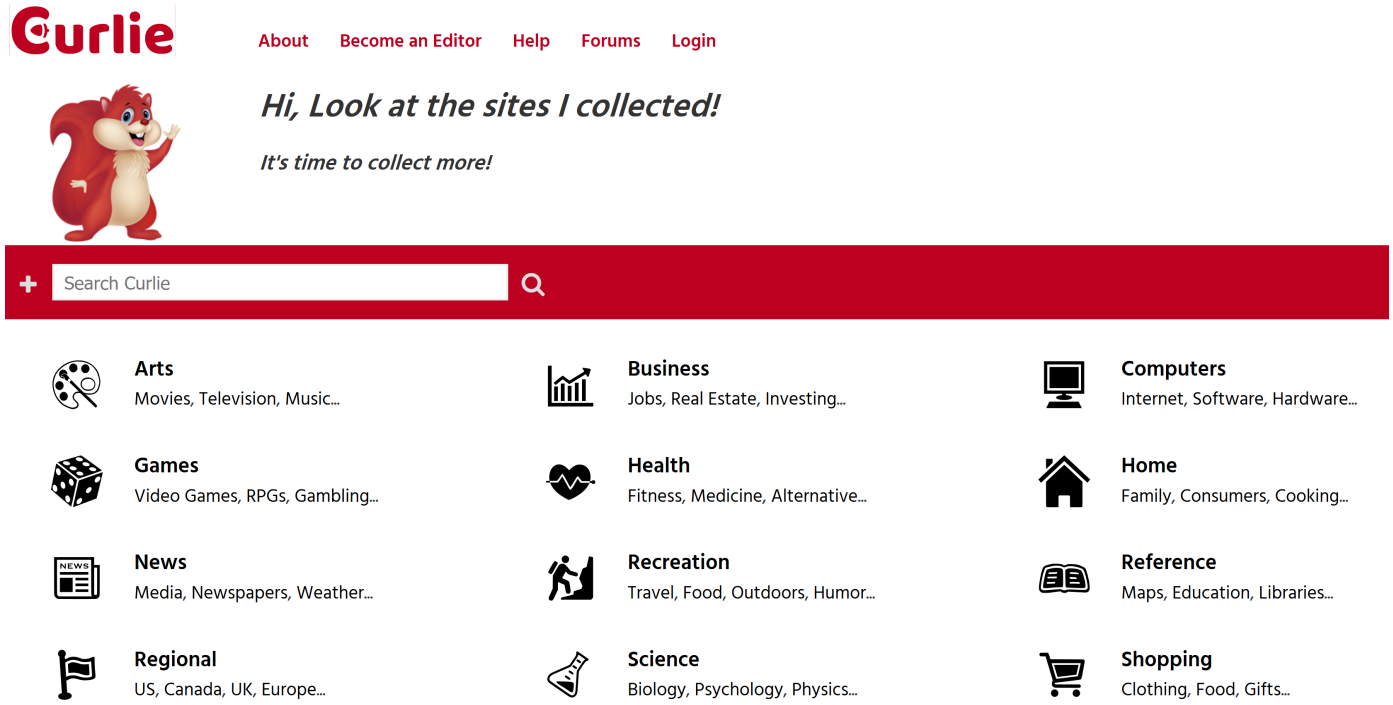
Curlie.org is the successor of the once-famous DMOZ directory of web pages.



Think about classifying web pages into the above categories. Why is it a misnomer to call this a “taxonomy”?

Classification

Curlie.org is the successor of the once-famous DMOZ directory of web pages.



Why is it a misnomer to call this a “taxonomy”?
Because there will be plenty of cases of dual classification.
In a real taxonomy, every item is classified into one and only category path.

Classification

Q: What are some other big (famous, influential) document taxonomies on the web?

Classification

Q: What are some other big (famous, influential) document taxonomies on the web?

A: IMDB, IAB, About.com, the Google Product Taxonomy, and many website taxonomies such as Amazon, CNET, etc.

Other Examples

Additional types of semantic analysis include:

- Tagging
- Topic segmentation
- Sentiment analysis

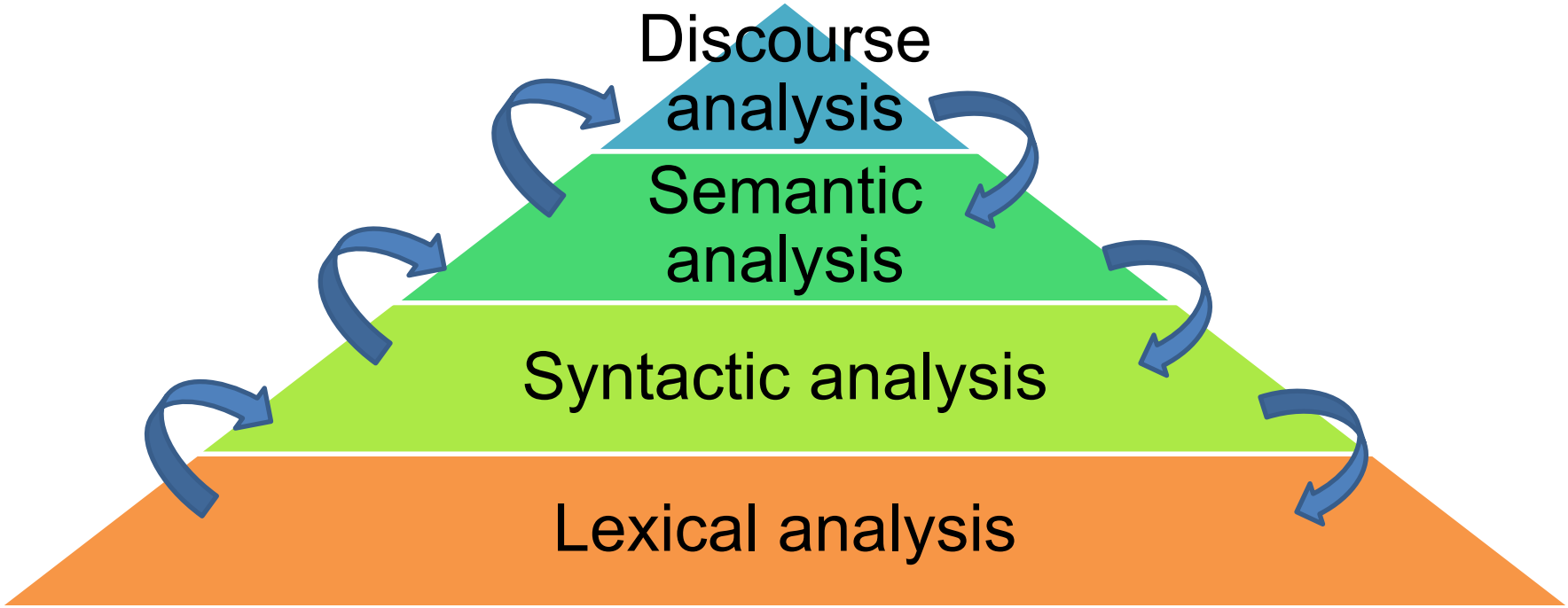
DataScience@SMU

Overview: Semantic Analysis in NLP

Natural Language Processing

Levels of Analysis

For convenience, we can think of there being higher and lower levels of analysis in NLP as long as we remember they are interconnected.



Semantic Analysis

Some major types:

- Named entity extraction (NEE a.k.a. NER)
- Relationship extraction (between NEs)
- Word sense disambiguation (WSD)
- Classification
- Tagging
- Topic segmentation
- Sentiment analysis

Named-Entity Extraction

- Also called “named-entity recognition” (abbreviated NEE or NER).
- Simple NEE just recognizes entities without typing.
- More sophisticated NEE does basic typing:
  - Persons
  - Organizations
  - Places
  - Events

Named-Entity Extraction

Good NEE will cluster together many variants, including epithets.

Hillary Clinton	Donald Trump
Hillary Rodham Clinton	The Donald
Senator Clinton	Donald J. Trump
Secretary Clinton	Donald Jackass Trump
Hitlery Clinton	Donald Drumpf
Crooked Hillary	Herr Trump

Relationship Extraction (between NEs)

- What relation does one NE have to another? Sometimes we can answer this during the NEE process.

These are sometimes called "triples" because there are three elements: two NEs and one relation.

Collections of triples are called "triple stores."

Word Sense Disambiguation (WSD)

WordNet Search - 3.1

- WordNet home page - Glossary - Help

Word to search for: chair Search WordNet

Display Options: (Select option to change) Change

Key: "S:" = Show Synset (semantic) relations, "W:" = Show Word (lexical) relations

Display options for sense: (gloss) "an example sentence"

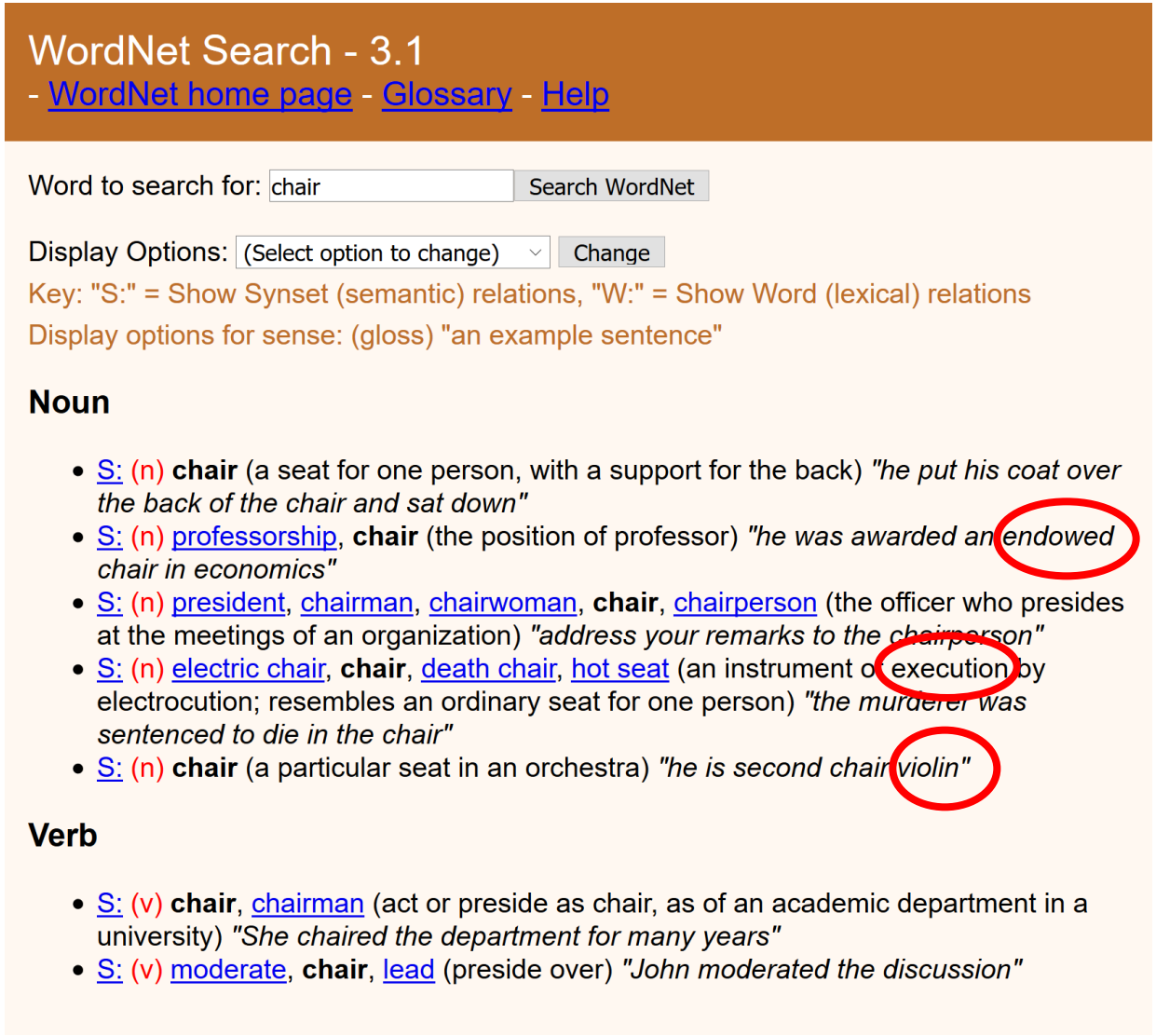
Noun

- S: (n) **chair** (a seat for one person, with a support for the back) *"he put his coat over the back of the chair and sat down"*
- S: (n) professorship, **chair** (the position of professor) *"he was awarded an endowed chair in economics"*
- S: (n) president, chairman, chairwoman, **chair**, chairperson (the officer who presides at the meetings of an organization) *"address your remarks to the chairperson"*
- S: (n) electric chair, **chair**, death chair, hot seat (an instrument of execution by electrocution; resembles an ordinary seat for one person) *"the murderer was sentenced to die in the chair"*
- S: (n) **chair** (a particular seat in an orchestra) *"he is second chair violin"*

Verb

- S: (v) **chair**, chairman (act or preside as chair, as of an academic department in a university) *"She chaired the department for many years"*
- S: (v) moderate, **chair**, lead (preside over) *"John moderated the discussion"*

Word Sense Disambiguation (WSD)



WordNet Search - 3.1
- WordNet home page - Glossary - Help

Word to search for: chair Search WordNet

Display Options: (Select option to change) Change

Key: "S:" = Show Synset (semantic) relations, "W:" = Show Word (lexical) relations

Display options for sense: (gloss) "an example sentence"

Noun

- S: (n) **chair** (a seat for one person, with a support for the back) *"he put his coat over the back of the chair and sat down"*
- S: (n) professorship, **chair** (the position of professor) *"he was awarded an endowed chair in economics"*
- S: (n) president, chairman, chairwoman, **chair**, chairperson (the officer who presides at the meetings of an organization) *"address your remarks to the chairperson"*
- S: (n) electric chair, **chair**, death chair, hot seat (an instrument of execution by electrocution; resembles an ordinary seat for one person) *"the murderer was sentenced to die in the chair"*
- S: (n) **chair** (a particular seat in an orchestra) *"he is second chair violin"*

Verb

- S: (v) **chair**, chairman (act or preside as chair, as of an academic department in a university) *"She chaired the department for many years"*
- S: (v) moderate, **chair**, lead (preside over) *"John moderated the discussion"*

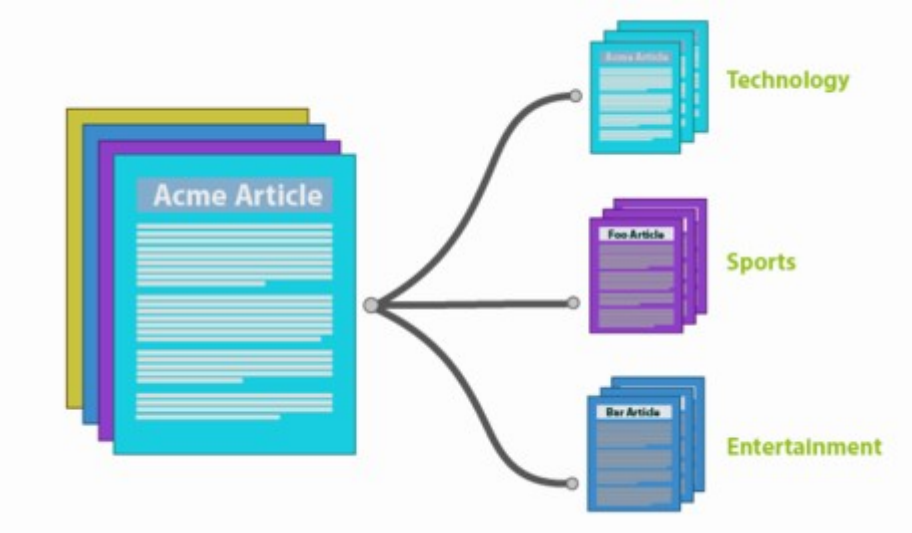
Context words give us a chance to disambiguate intended senses.

Doing this at scale is hard! It's still an open research area of NLP (i.e., unsolved).

Classification

- In classification, we use a tree-structured graph to place documents into categories.
- Often we employ machine learning methods, such as SVM*.

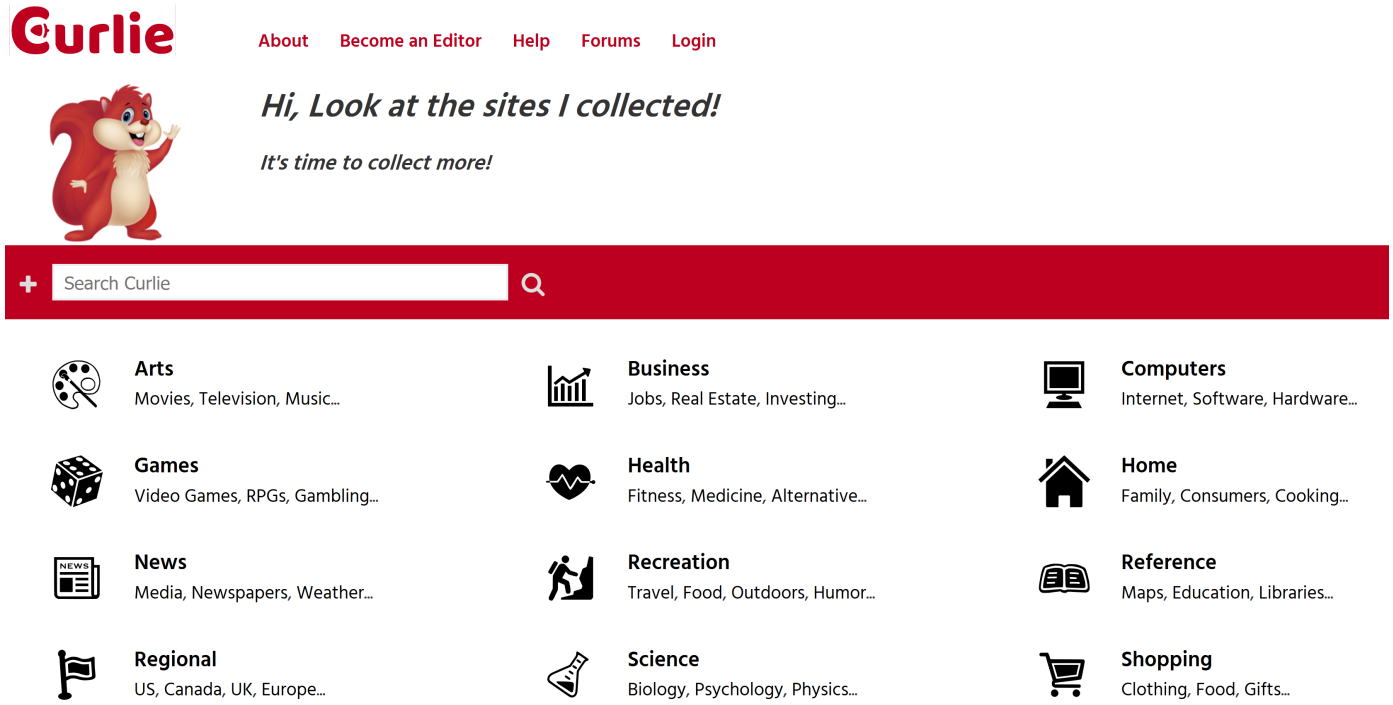
We always start from a preexisting category schema, often called a "taxonomy."



*support vector machine

Classification

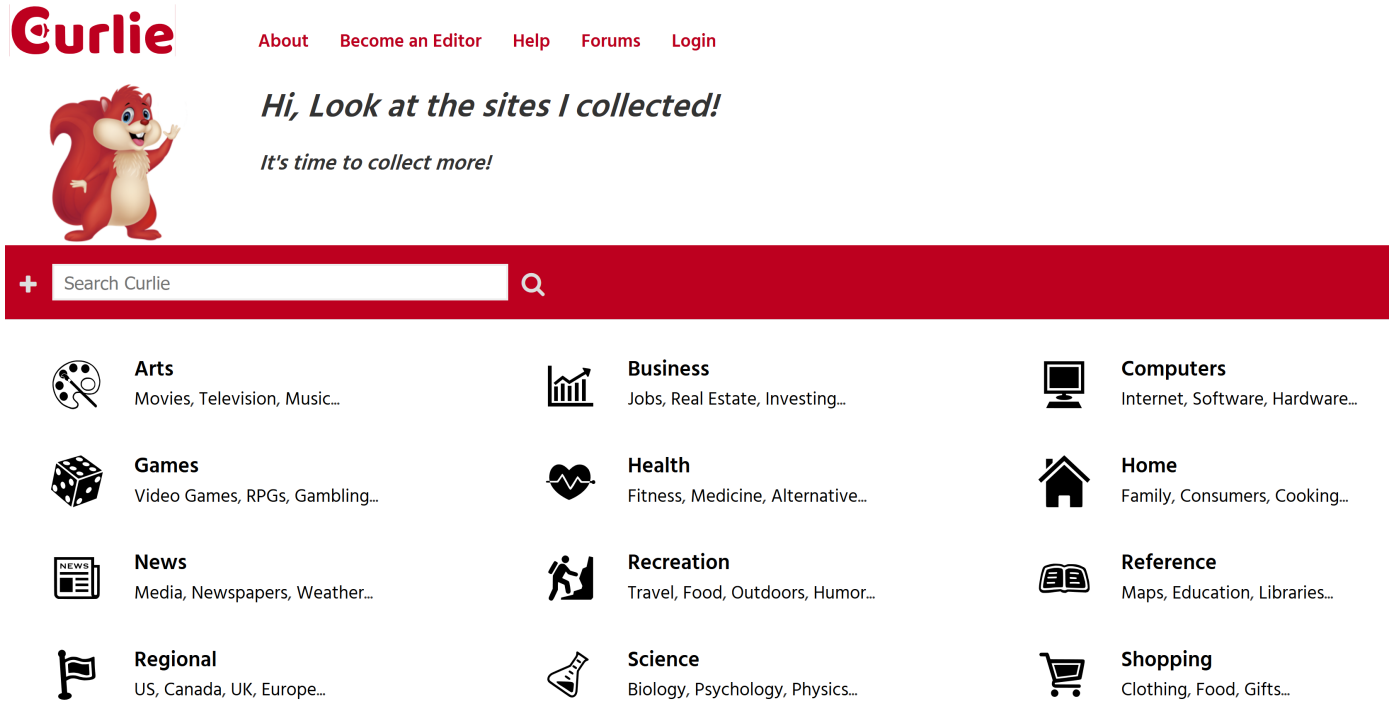
Curlie.org is the successor of the once-famous DMOZ directory of web pages.



Think about classifying web pages into the above categories. Why is it a misnomer to call this a “taxonomy”?

Classification

Curlie.org is the successor of the once-famous DMOZ directory of web pages.



Why is it a misnomer to call this a “taxonomy”?
Because there will be plenty of cases of dual classification.
In a real taxonomy, every item is classified into one and only category path.

Classification

Q: What are some other big (famous, influential) document taxonomies on the web?

Classification

Q: What are some other big (famous, influential) document taxonomies on the web?

A: IMDB, IAB, About.com, the Google Product Taxonomy, and many website taxonomies such as Amazon, CNET, etc.

Other Examples

Additional types of semantic analysis include:

- Tagging
- Topic segmentation
- Sentiment analysis

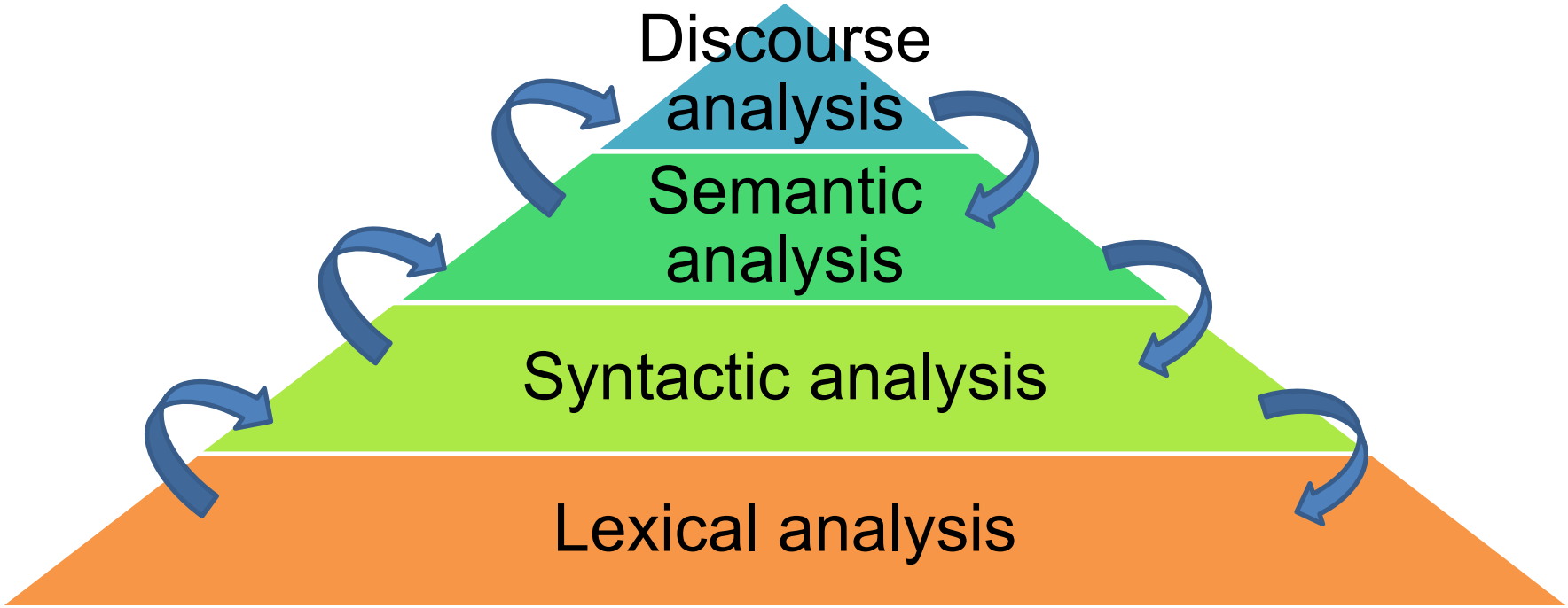
DataScience@SMU

Overview: Discourse Analysis in NLP

Natural Language Processing

Levels of Analysis

For convenience, we can think of there being higher and lower levels of analysis in NLP as long as we remember they are interconnected.



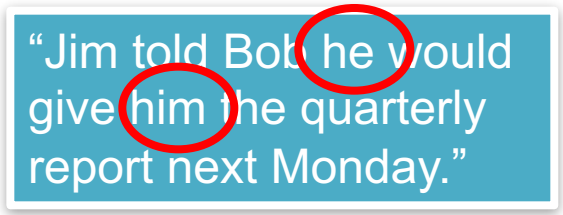
Discourse Analysis

Some major types:

- Anaphora resolution
- Discourse modeling
- Question answering
- Textual entailment
- Pragmatic analysis

Anaphora Resolution

- A pronoun always refers to its most recent antecedent, right?
- No. Pronouns don't always behave that way



Who is "he" and who is "him"?

Discourse Modeling

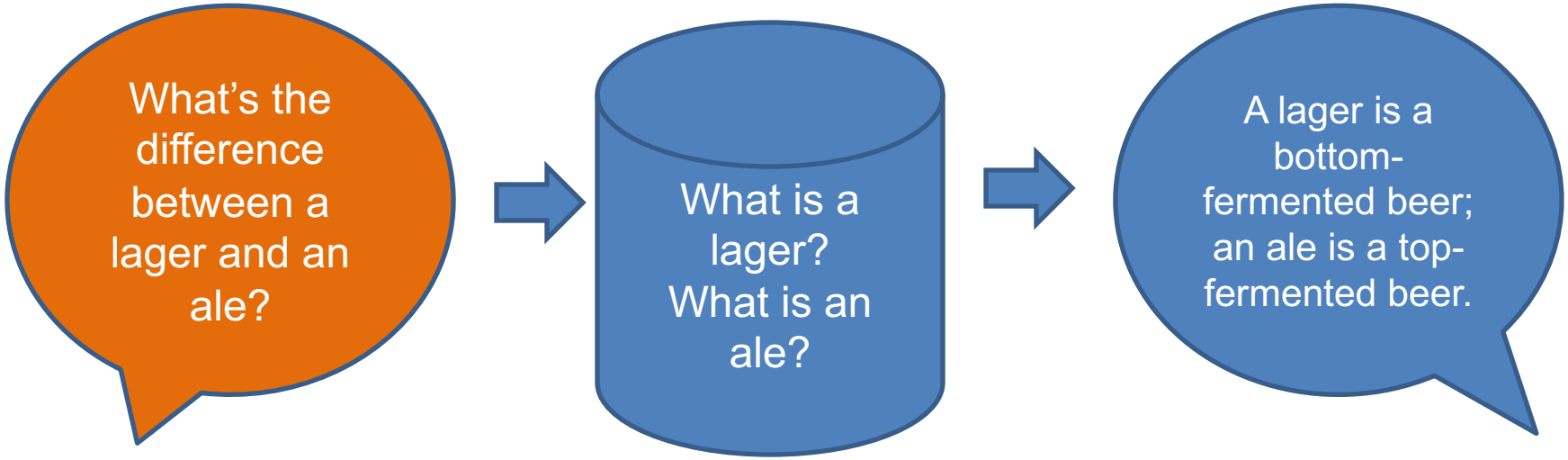
Waiter: Introduce.Self.Name
Waiter: Introduce.Self.Role
Waiter: Offer.Beverage
Customer: Order.Beverage
Waiter: Offer.Appetizer
Customer: Order.Appetizer
Etc.

Roger Schank's work on "scripts" in the '70s…

…was anticipated by Claude Lévi-Strauss's "structuralism" in the '40s.

Question Answering

- Most straightforward approach is matching preexisting Q-A materials
- The matching is often not straightforward. We can build models that adapt question formats.



We may not have the exact question in our db but can combine other questions (with answers) to make a response.

Applying Inference to Text

- Textual entailment

"All men are mortal."
"Socrates is a man."

Inference: Socrates is mortal.

This inference follows from strict logic.

- Pragmatic analysis

"Did you like the play?"
"It was, um, interesting."

Inference: He didn't like the play.

This inference ***doesn't*** follow from strict logic.

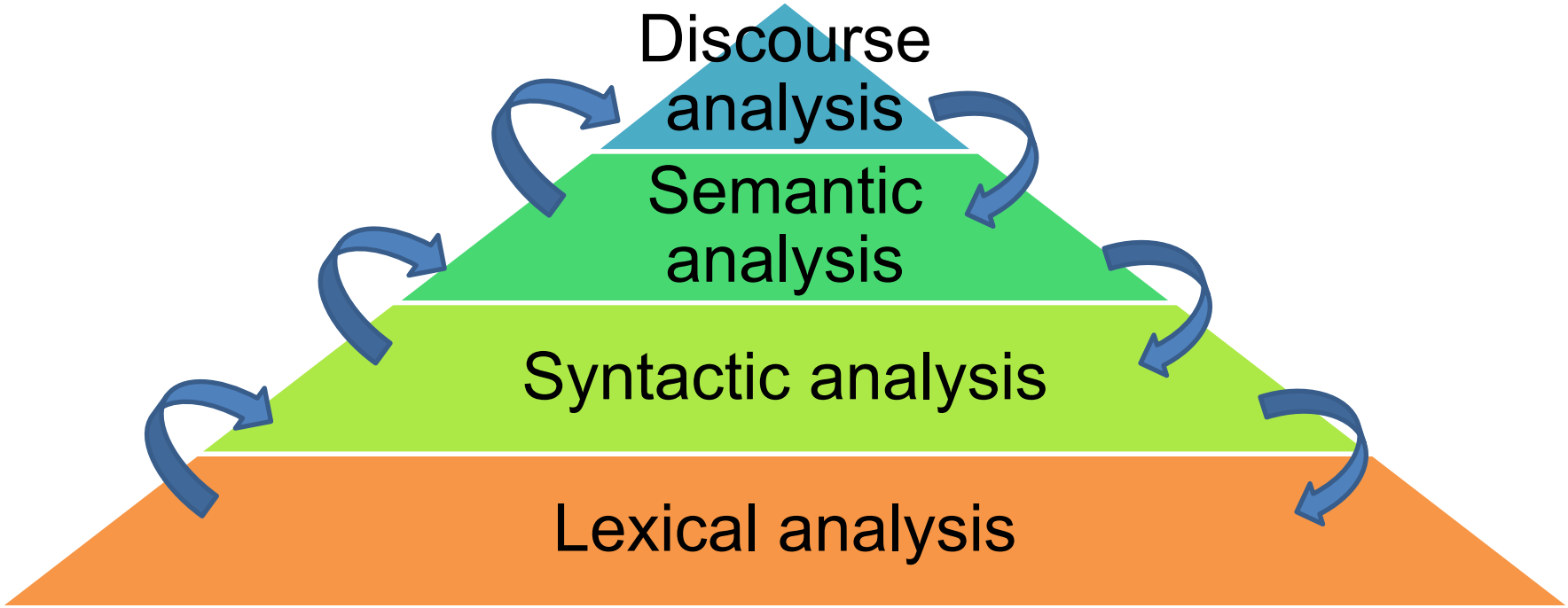
DataScience@SMU

Review: Levels of Analysis and Applications in NLP

Natural Language Processing

Levels of Analysis

For convenience, we can think of there being higher and lower levels of analysis in NLP as long as we remember they are interconnected.



Levels of Analysis and Implementations

Generally, we use lower levels also when we move to “higher” levels of analysis.

This table just shows *typical* levels of analysis employed in NLP.

One can often enhance results by adding a higher level than is strictly necessary, e.g., adding syntactic analysis will improve spell correction.

***Minimum* level of analysis generally used for implementations**	**Lexical Analysis**	**Syntactic Analysis**	**Semantic Analysis**	**Discourse Analysis**
Stemming	X			
Keyword Tagging	X			
Spell Correction	X			
Terminology Extraction	X			
Reading Level Estimation	X			
Sentence-Boundary Detection	X	X		
POS-Tagging	X	X		
Parsing	X	X		
Lemmatization	X	X		
Unitizing/Normalizing	X	X		
NEE	X	X	X	
WSD	X	X	X	
Conceptual Tagging	X	X	X	
Classification	X	X	X	
Topic Segmentation	X	X	X	
Sentiment Analysis	X	X	X	
Anaphora Resolution	X	X	X	X
Discourse Modeling	X	X	X	X
Question Answering	X	X	X	X
Textual Entailment	X	X	X	X
Pragmatic Analysis	X	X	X	X

DataScience@SMU