CS114 Lecture 22

Discourse Structure

April 28, 2014

Professor Meteer

Thanks for Jurafsky & Martin & Prof. Pustejovksy for slides

Outline

- Discourse Structure
  - Textiling
- Coherence
  - Hobbs coherence relations
  - Rhetorical Structure Theory
- Coreference
  - Kinds of reference phenomena
  - Constraints on co-reference
  - Anaphora Resolution
    - Hobbs
    - Loglinear
  - Coreference

Part I: Discourse Structure

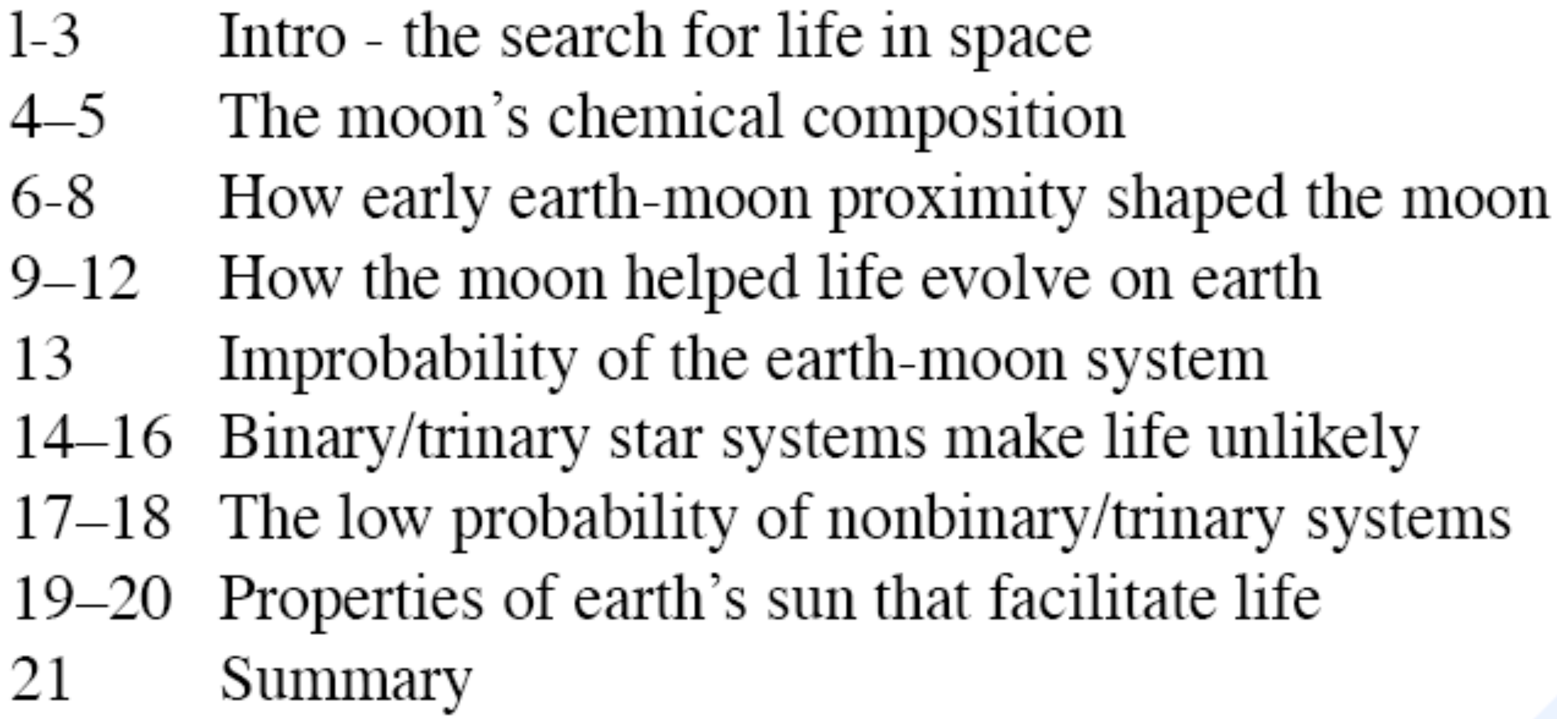
- Conventional structures for different genres
  - Academic articles:
    - Abstract, Introduction, Methodology, Results, Conclusion
  - Newspaper story:
    - inverted pyramid structure (**lead** followed by expansion)

Discourse Segmentation

- Simpler task
  - Discourse segmentation
    - Separating document into linear sequence of subtopics
- Why?
  - Information retrieval:
    - automatically segmenting a TV news broadcast or a long news story into sequence of stories
  - Text summarization:
    - summarize different segments of a document
  - Information extraction:
    - Extract info from inside a single discourse segment

Unsupervised Discourse Segmentation

- Hearst (1997): 21-paragraph science news article called “Stargazers”
- Goal: produce the following subtopic segments:



1-3	Intro - the search for life in space
4–5	The moon’s chemical composition
6-8	How early earth-moon proximity shaped the moon
9–12	How the moon helped life evolve on earth
13	Improbability of the earth-moon system
14–16	Binary/trinary star systems make life unlikely
17–18	The low probability of nonbinary/trinary systems
19–20	Properties of earth’s sun that facilitate life
21	Summary

Key intuition: **cohesion**

- Halliday and Hasan (1976): “The use of certain linguistic devices to link or tie together textual units”

- Lexical cohesion:
  - Indicated by relations between words in the two units (**identical word**, **synonym**, **hypernym**)
    - Before winter **I** built a chimney, and **shingled** the sides of my **house.**

      I thus have a tight **shingled** and plastered **house**.

    - Peel, core and slice **the pears and the apples**. Add **the fruit** to the skillet.

Key intuition: **cohesion**

- Non-lexical: anaphora
  - The **Woodhouses** were first in consequence there. All looked up to **them**.
- Cohesion chain:
  - Peel, core and slice **the pears and the apples**. Add **the fruit** to the skillet. When **they** are soft…

Intuition of cohesion-based segmentation

- Sentences or paragraphs in a subtopic are cohesive with each other

- But not with paragraphs in a neighboring subtopic

- Thus if we measured the cohesion between every neighboring sentences
  - We might expect a ‘dip’ in cohesion at subtopic boundaries.

What makes a text/dialogue coherent?

"Consider, for example, the difference between passages (18.71) and (18.72). Almost certainly not. The reason is that these utterances, when juxtaposed, will not exhibit coherence. Do you have a discourse? Assume that you have collected an arbitrary set of well-formed and independently interpretable utterances, for instance, by randomly selecting one sentence from each of the previous chapters of this book."

vs….

Coherence regained

"Assume that you have collected an arbitrary set of well-formed and independently interpretable utterances, for instance, by randomly selecting one sentence from each of the previous chapters of this book. Consider, for example, the difference between passages (18.71) and (18.72). Do you have a discourse? Almost certainly not. The reason is that these utterances, when juxtaposed, will not exhibit coherence."

What makes a text coherent?

- Discourse/topic structure
  - Appropriate sequencing of subparts of the discourse
- Rhetorical structure
  - Appropriate use of coherence relations between subparts of the discourse
- Referring expressions
  - Words or phrases, the *semantic interpretation* of which is *a discourse entity*

Information Status

- Contrast
  - John wanted a *poodle* but Becky preferred a *corgi*.
- Topic/comment
  - *The corgi they bought* turned out to have fleas.
- Theme/rheme
  - *The corgi they bought* turned out to have fleas.
- Focus/presupposition
  - It was Becky who took him to the vet.
- Given/new
  - Some wildcats bite, but this wildcat turned out to be a sweetheart.
  - Contrast Speaker (S) and Hearer (H)

Determining Given vs. New

- Entities when first introduced are new
  - Brand-new (H must create a new entity)

    I saw a dinosaur today.
  - Unused (H already knows of this entity)

    I saw your mother today.
- Evoked entities are old -- already in the discourse
  - Textually evoked

    The dinosaur was scaley and gray.
  - Situationally evoked

    The light was red when you went through it.
- Inferrables
  - Containing

    I bought a carton of eggs. One of them was broken.
  - Non-containing

    A bus pulled up beside me. The driver was a monkey.

Given/New and Definiteness/Indefiniteness

- Subject NPs tend to be syntactically *definite* and *old*
- Object NPs tend to be *indefinite* and *new*

  I saw a black cat yesterday. *The cat* looked hungry.

  - Definite articles, demonstratives, possessives, personal pronouns, proper nouns, quantifiers like all, every
- Indefinite articles, quantifiers like some, any, one signal indefiniteness…but….

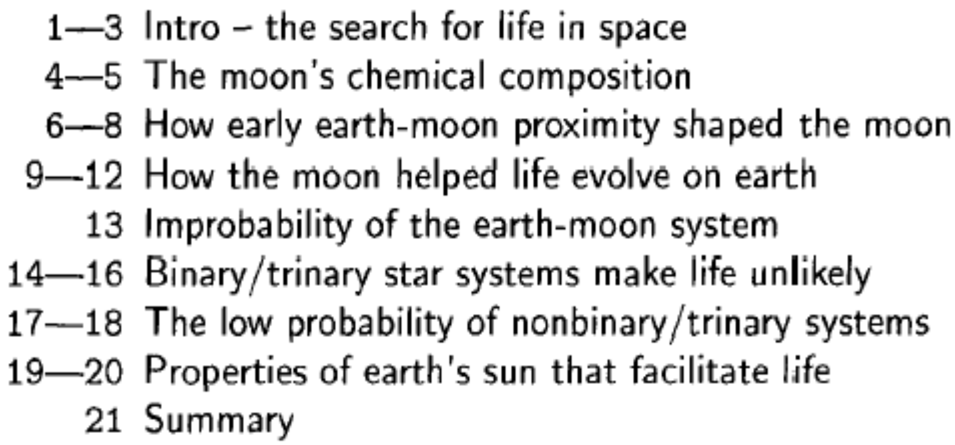
  This guy came into the room

Discourse/Topic Structure

- Text Segmentation:
  - Linear
    - TextTiling
    - Look for changes in content words
  - Hierarchical
    - Grosz & Sidner’s Centering theory
    - Morris & Hirst’s algorithm
    - Lexical chaining through Roget’s thesaurus
  - Hierarchical + Relations
    - Mann et al.’s Rhetorical Structure Theory
    - Marcu’s algorithm

TextTiling (Hearst 94)

- Goal: find multi-paragraph topics
- Example: 21 paragraph article called *Stargazers*

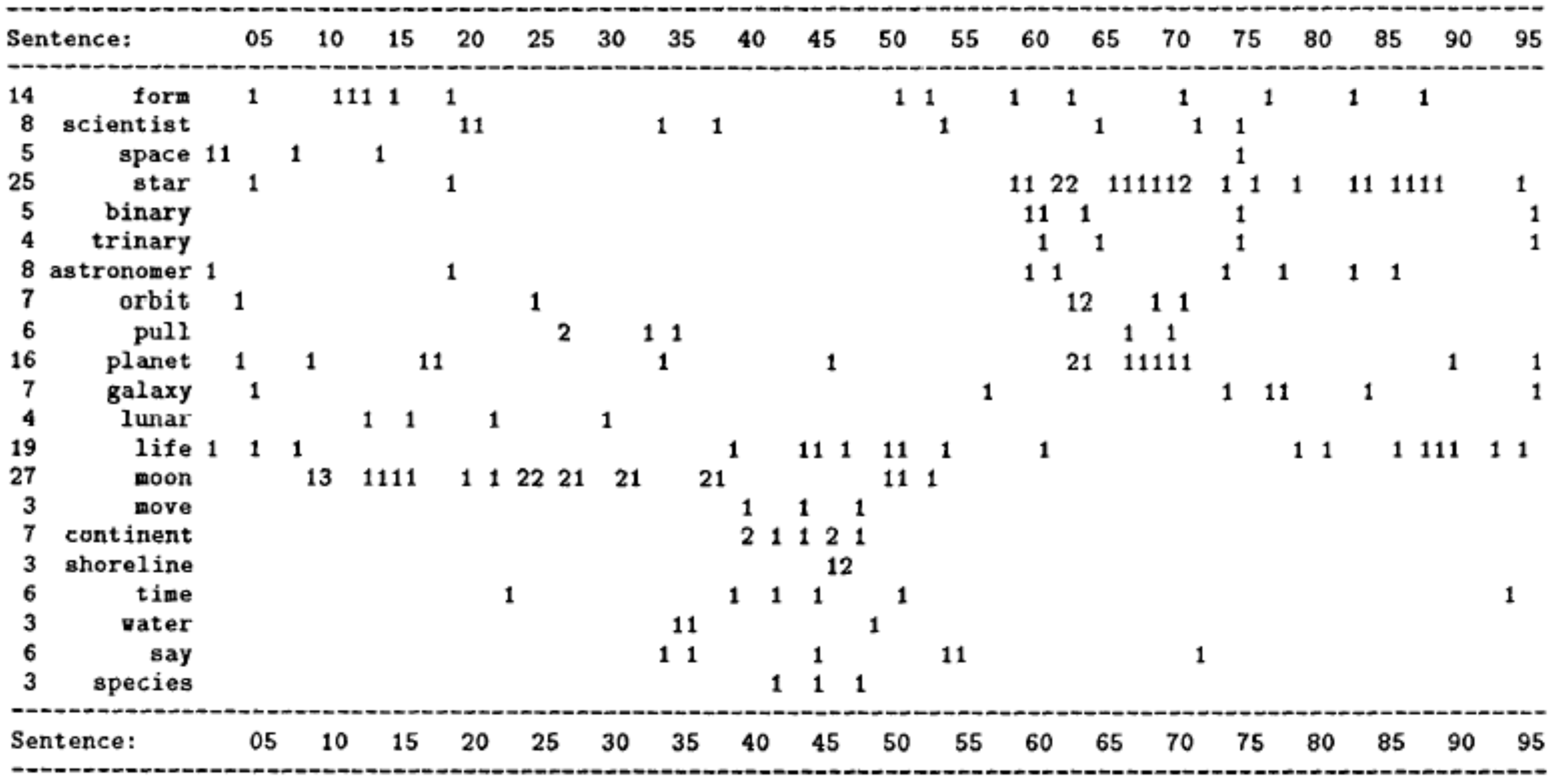


1—3	Intro – the search for life in space
4—5	The moon's chemical composition
6—8	How early earth-moon proximity shaped the moon
9—12	How the moon helped life evolve on earth
13	Improbability of the earth-moon system
14—16	Binary/trinary star systems make life unlikely
17—18	The low probability of nonbinary/trinary systems
19—20	Properties of earth's sun that facilitate life
21	Summary

TextTiling (Hearst 94)

- Goal: find multi-paragraph topics
- But … it's difficult to define topic (Brown & Yule)
- Focus instead on *topic shift* or *change*
- Change in *content*, by contrast with setting, scene, characters
- Mechanism:
  - compare adjacent blocks of text
  - look for shifts in vocabulary

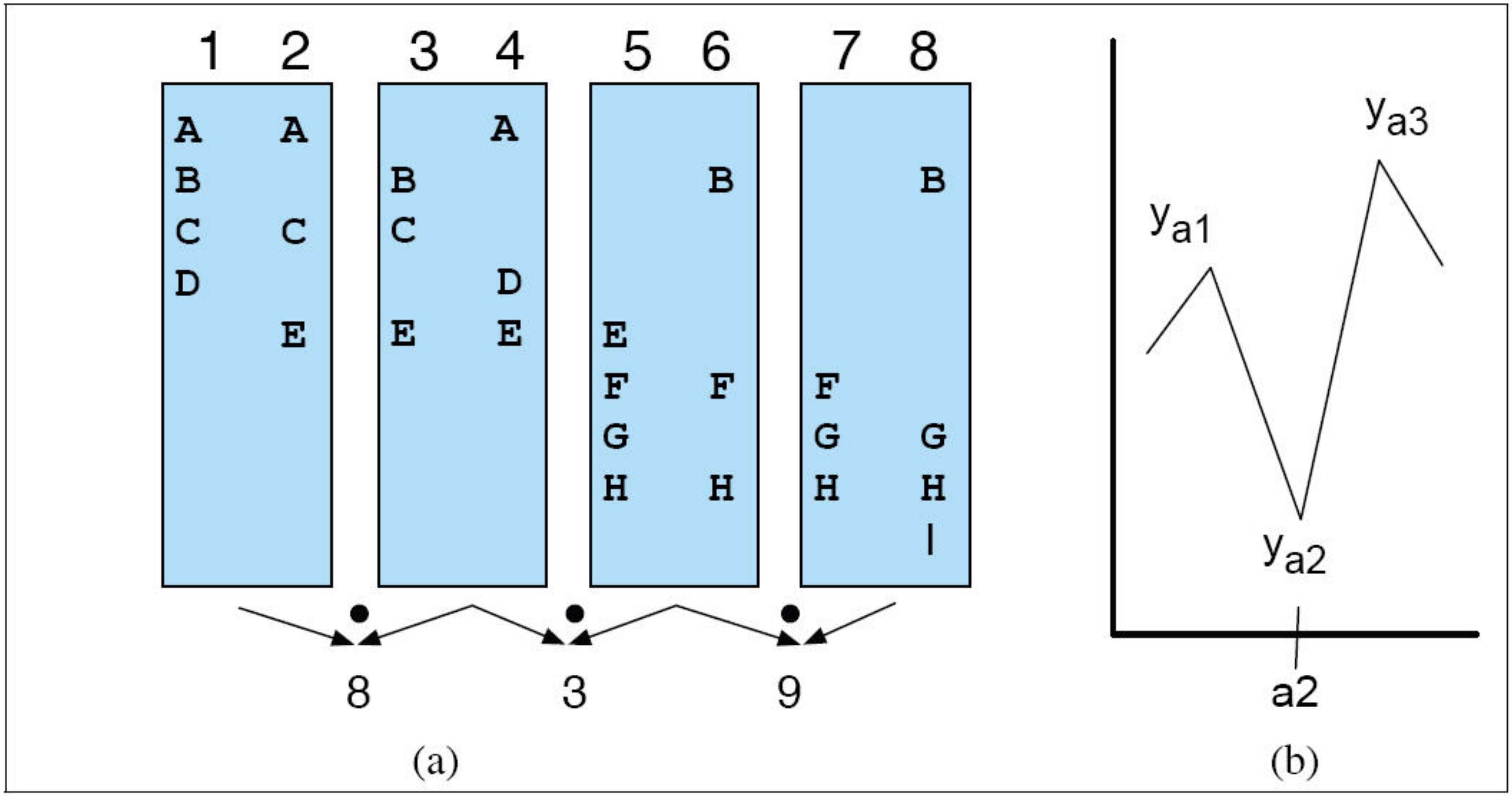
Intuition behind TextTiling



```
-------------------------------------------------------------------------------------------------------------
Sentence:        05   10   15   20   25   30   35   40   45   50   55   60   65   70   75   80   85   90   95
-------------------------------------------------------------------------------------------------------------
14       form    1     111 1   1                               1 1     1   1       1     1     1    1
 8  scientist                   11            1   1               1          1      1  1
 5      space 11    1     1                                                            1
25       star    1             1                                       11 22  111112  1 1  1   11 1111     1
 5     binary                                                            11  1          1                    1
 4    trinary                                                             1   1         1                    1
 8 astronomer 1                1                                        1 1           1   1    1  1
 7      orbit   1                    1                                     12    1 1
 6       pull                          2     1 1                               1  1
16     planet   1    1       11               1           1                21  11111                  1     1
 7     galaxy    1                                                   1                1  11     1           1
 4      lunar            1  1     1       1
19       life 1  1  1                              1    11 1  11  1      1                 1 1    1 111  1 1
27       moon        13  1111   1 1 22 21  21    21           11 1
 3       move                                        1   1   1
 7  continent                                        2 1 1 2 1
 3  shoreline                                              12
 6       time                      1               1  1  1     1                                          1
 3      water                                  11            1
 6        say                                 1 1        1        11                1
 3    species                                          1  1  1
-------------------------------------------------------------------------------------------------------------
Sentence:        05   10   15   20   25   30   35   40   45   50   55   60   65   70   75   80   85   90   95
-------------------------------------------------------------------------------------------------------------
```

Figure 2: Distribution of selected terms from the *Stargazer* text, with a single digit frequency per sentence number (blanks indicate a frequency of zero).

Figure 21.1


1
2
3
4
5
6
7
8
A
A
A
B
B
B
B
C
C
C
D
D
E
E
E
E
F
F
F
G
G
G
H
H
H
H
I
8
3
9
(a)
y_{a1}
y_{a3}
y_{a2}
a2
(b)

TextTiling Algorithm

- Tokenization
- Lexical Score Determination
  - Blocks
  - Vocabulary Introductions
  - Chains
- Boundary Identification

Tokenization

- Convert text stream into terms (words)
- Remove “stop words”
- Reduce to root (inflectional morphology)
- Subdivide into “token-sequences”
  - (substitute for sentences)
- Find potential boundary points
  - (paragraphs breaks)

Determining Scores

- Compute a score at each token-sequence gap
- Score based on lexical occurrences
- Block algorithm:

$$score(i) = \frac{\sum_t w_{t,b_1} w_{t,b_2}}{\sqrt{\sum_t w_{t,b_1}^2 \sum_t w_{t,b_2}^2}}$$

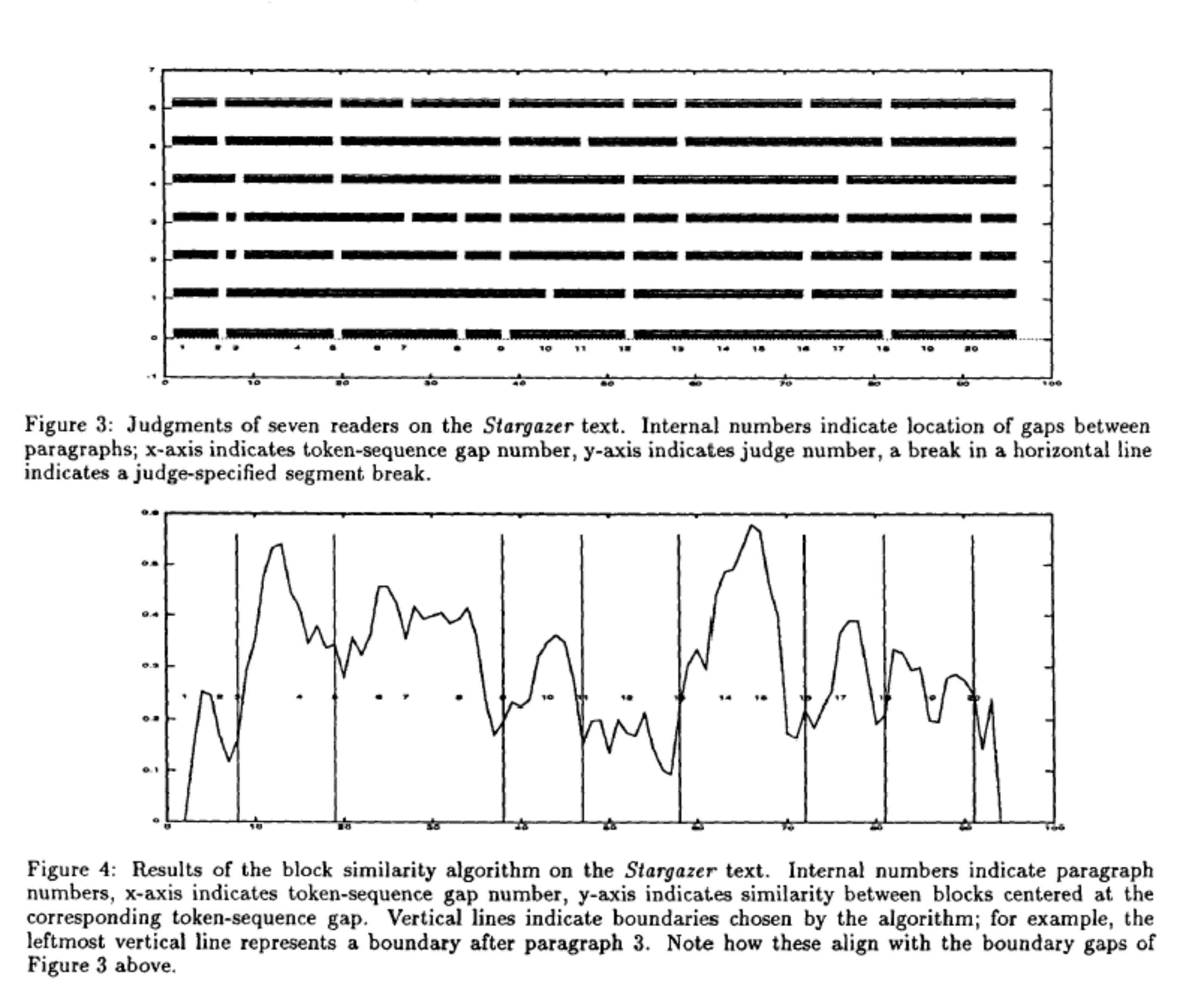


Figure 3: Judgments of seven readers on the *Stargazer* text. Internal numbers indicate location of gaps between paragraphs; x-axis indicates token-sequence gap number, y-axis indicates judge number, a break in a horizontal line indicates a judge-specified segment break.

Figure 4: Results of the block similarity algorithm on the *Stargazer* text. Internal numbers indicate paragraph numbers, x-axis indicates token-sequence gap number, y-axis indicates similarity between blocks centered at the corresponding token-sequence gap. Vertical lines indicate boundaries chosen by the algorithm; for example, the leftmost vertical line represents a boundary after paragraph 3. Note how these align with the boundary gaps of Figure 3 above.

Boundary Identification

- Smooth the plot (average smoothing)
- Assign depth score at each token-sequence gap
- “Deeper” valleys score higher
- Order boundaries by depth score
- Choose boundary cut off (avg-sd/2)

Evaluation

- Data
  - Twelve news articles from Dialog
  - Seven human judges per article
  - “major” boundaries: chosen by >= 3 judges
  - Avg number of paragraphs: 26.75
  - Avg number of boundaries: 10 (39%)

- Results
  - Between upper and lower bounds
  - Upper bound: judges’ averages
  - Lower bound: reasonable simple algorithm

Assessing Agreement Among Judges

***KAPPA* Coefficient**

- **Measures pairwise agreement**
- **Takes *expected chance* agreement into account**
- **P(A) = proportion of times judges agree**
- **P(E) = proportion expected chance agreement**

$$k = \frac{P(A) - P(E)}{1 - P(E)}$$

- **.43 to .68 (Isard & Carletta 95, boundaries)**
- **.65 to .90 (Rose 95, sentence segmentation)**
- **Here, *k*= .647**

$$P(E) = (P(B))^2 + (1 - P(B))^2$$

$$P(E) = .39^2 + .61^2 = .52$$

Part II: Text Coherence

What makes a discourse coherent?

The reason is that these utterances, when juxtaposed, will not exhibit coherence. Almost certainly not. Do you have a discourse? Assume that you have collected an arbitrary set of well-formed and independently interpretable utterances, for instance, by randomly selecting one sentence from each of the previous chapters of this book.

Better?

Assume that you have collected an arbitrary set of well-formed and independently interpretable utterances, for instance, by randomly selecting one sentence from each of the previous chapters of this book. Do you have a discourse? Almost certainly not. The reason is that these utterances, when juxtaposed, will not exhibit coherence.

Coherence

- John hid Bill’s car keys. He was drunk.

- ??John hid Bill’s car keys. He likes spinach.

What makes a text coherent?

- Appropriate use of coherence relations between subparts of the discourse -- rhetorical structure

- Appropriate sequencing of subparts of the discourse -- discourse/topic structure

- Appropriate use of referring expressions

Hobbs 1979 Coherence Relations

- “ Result ” :
  - Infer that the state or event asserted by S0 causes or could cause the state or event asserted by S1.
    - The Tin Woodman was caught in the rain. His joints rusted.

Hobbs: “Explanation”

- Infer that the state or event asserted by S1 causes or could cause the state or event asserted by S0.
  - John hid Bill’s car keys. He was drunk.

Hobbs: “Parallel”

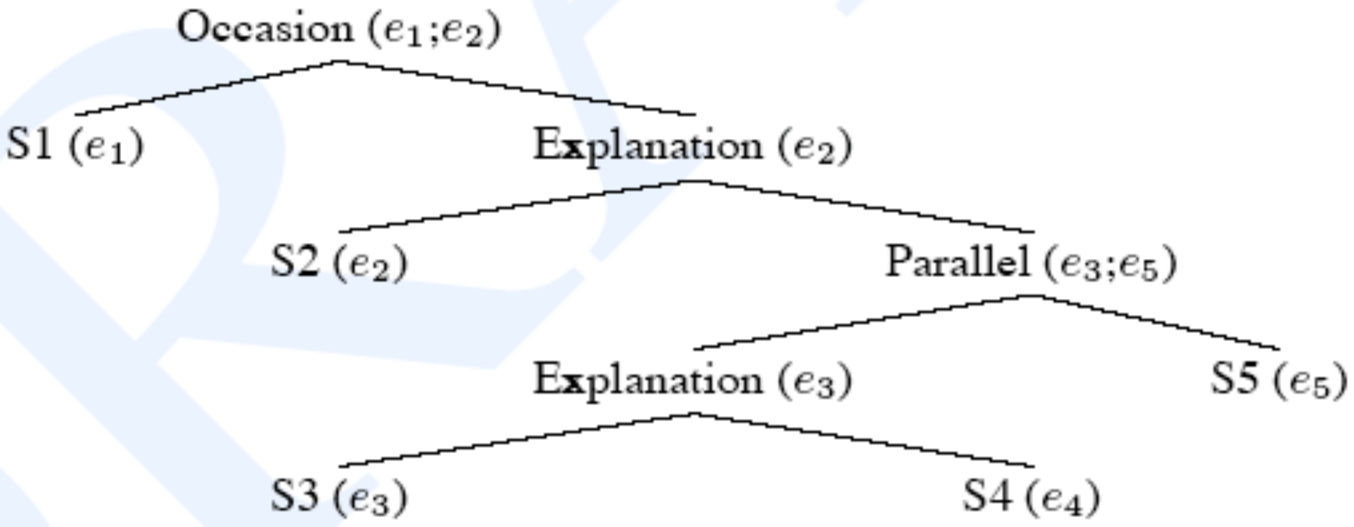
- Infer p(a1, a2..) from the assertion of S0 and p(b1,b2…) from the assertion of S1, where ai and bi are similar, for all I.
  - The Scarecrow wanted some brains. The Tin Woodman wanted a heart.

Hobbs “Elaboration”

- Infer the same proposition P from the assertions of S0 and S1.
  - Dorothy was from Kansas. She lived in the midst of the great Kansas prairies.

Coherence relations impose a discourse structure

John went to the bank to deposit his paycheck. (S1)
He then took a train to Bill's car dealership. (S2)
He needed to buy a car. (S3)
The company he works for now isn't near any public transportation. (S4)
He also wanted to talk to Bill about their softball league. (S5)



Occasion (e_1;e_2)

S1 (e_1)

Explanation (e_2)

S2 (e_2)

Parallel (e_3;e_5)

Explanation (e_3)

S5 (e_5)

S3 (e_3)

S4 (e_4)

Rhetorical Structure Theory

- Another theory of discourse structure, based on identifying relations between segments of the text

  - Nucleus/satellite notion encodes asymmetry
    - Nucleus is thing that if you deleted it, text wouldn't make sense.

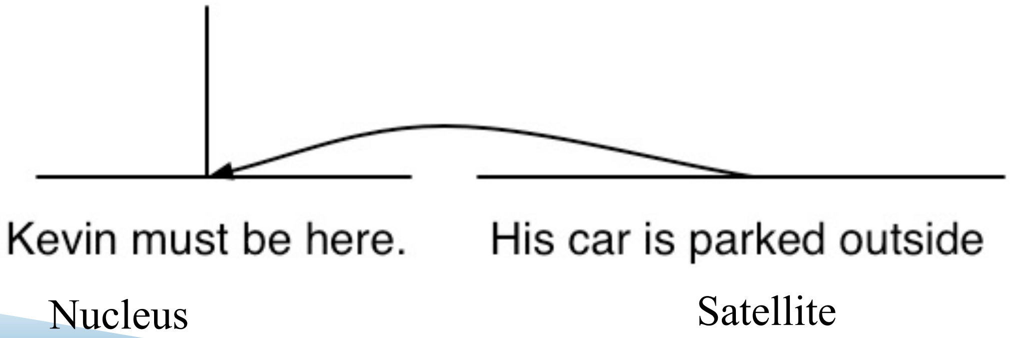
  - Some rhetorical relations:
    - Elaboration: (set/member, class/instance, whole/part…)
    - Contrast: multinuclear
    - Condition: Sat presents precondition for N
    - Purpose: Sat presents goal of the activity in N

One example of rhetorical relation

- A sample definition
  - Relation: Evidence
  - Constraints on N: H might not believe N as much as S think s/he should
  - Constraints on Sat: H *already believes or will believe* Sat
  - Effect: H's belief in N is increased
- An example:

  Kevin must be here.

  His car is parked outside.



Automatic Rhetorical Structure Labeling

- Supervised machine learning
  - Get a group of annotators to assign a set of RST relations to a text
  - Extract a set of surface features from the text that might signal the presence of the rhetorical relations in that text
  - Train a supervised ML system based on the training set

Features: **cue phrases**

- Explicit markers: *because, however, therefore, then, etc.*
- Tendency of certain syntactic structures to signal certain relations:

  Infinitives are often used to signal purpose relations: Use rm *to delete files.*
- Ordering
- Tense/aspect
- Intonation

Some Problems with RST

- How many Rhetorical Relations are there?
- How can we use RST in dialogue as well as monologue?
- RST does not model overall structure of the discourse.
- Difficult to get annotators to agree on labeling the same texts

Dialogue acts

- Also called “conversational moves”
- An act with (internal) structure related specifically to its dialogue function
- Incorporates ideas of grounding
- Incorporates other dialogue and conversational functions that Austin and Searle didn’t seem interested in

Slides from Dan Jurafsky and Paul Martin

Verbmobil Dialogue Acts

THANK	thanks
GREET	Hello Dan
INTRODUCE	It's me again
BYE	Allright, bye
REQUEST-COMMENT	How does that look?
SUGGEST	June 13th through 17th
REJECT	No, Friday I'm booked all day
ACCEPT	Saturday sounds fine
REQUEST-SUGGEST	What is a good day of the week for you?
INIT	I wanted to make an appointment with you
GIVE_REASON	Because I have meetings all afternoon
FEEDBACK	Okay
DELIBERATE	Let me check my calendar here
CONFIRM	Okay, that would be wonderful
CLARIFY	Okay, do you mean Tuesday the 23rd?

Slides from Dan Jurafsky and Paul Martin

DAMSL (Dialog Act Markup in Several Layers

- DAMSL distinguishes four dimensions according to the unit's purpose and role in dialogue:
  - Communicative status: whether utterance is intelligible and whether it was successfully completed (uninterpretable, abandoned, self-talk)
  - Information level: abstract characterization of the semantic content
    - Task: utterances that advance the task
    - Task-management: utterances that discuss the problem solving process or experimental scenario
    - Communication management:,conventional phrases that maintain contact, perception, and under- standing during the communication process: greetings, closings, acknowledgements ("Okay", "uh-huh"), stalling for time ("Okay", "Let me see"), signals of speech repairs ("oops") or misunderstandings ("sorry?", "huh?")
    - Other-level
  - Forward-looking function: characterizes what effect an utterance has on subsequent dialogue and interaction
  - Backward-looking function: captures the way the current utterance is related to the previous dialogue

DAMSL: forward looking func.

STATEMENT	a claim made by the speaker
INFO-REQUEST	a question by the speaker
CHECK	a question for confirming information
INFLUENCE-ON-ADDRESSEE	(=Searle's directives)
OPEN-OPTION	a weak suggestion or listing of options
ACTION-DIRECTIVE	an actual command
INFLUENCE-ON-SPEAKER	(=Austin's commissives)
OFFER	speaker offers to do something
COMMIT	speaker is committed to doing something
CONVENTIONAL	other
OPENING	greetings
CLOSING	farewells
THANKING	thanking and responding to thanks

Slides from Dan Jurafsky and Paul Martin

Forward looking (2006)

- Statement:
  - Asserts and other acts where the speaker makes a claim about the world (modified in Core et al., 1998 to also allow statements to be claims about the communication).
- Info-request:
  - Speaker requests Hearer (by just asking or in another, indirect way) to provide information.
- Influencing-addressee-future-action:
  - Speaker is suggesting potential action to Hearer, beyond answering a request for information.
- Committing-speaker-future-action:
  - Speaker is potentially committing himself to perform a future action.
- Conventional:
  - Opening or Closing, i.e. Speaker summons Hearer and/or starts the interaction, or Speaker closes the dialogue or is dismissing Hearer.
- Explicit-performative:
  - Speaker is performing an action by virtue of making the utterance.
- Exclamation (no explicit definition given)
- Other-forward-looking-function:
  - No definition given; supposedly any FLF that does not fit into the categories 1-7.

DAMSL: backward looking func.

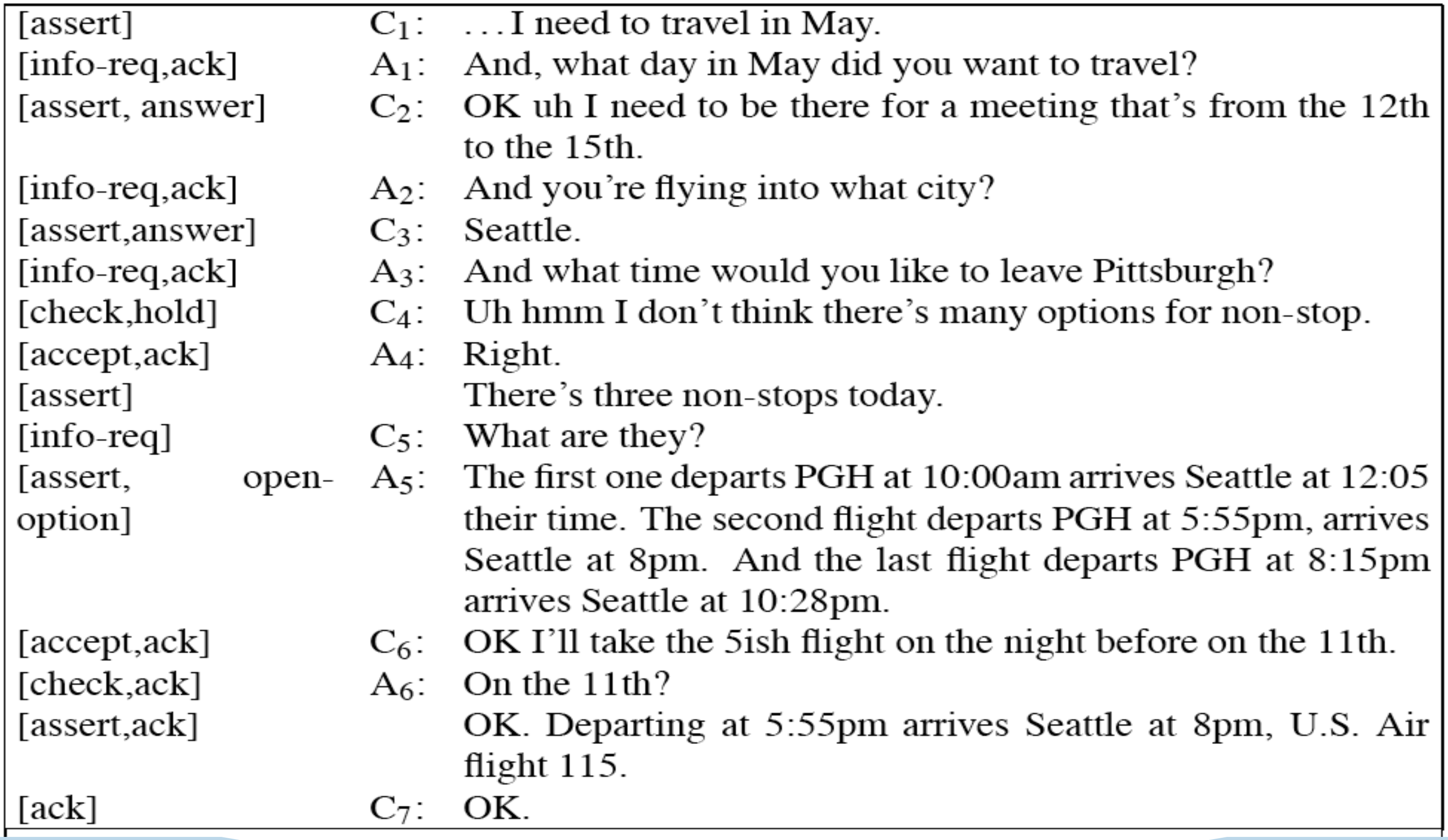
AGREEMENT	speaker's response to previous proposal
ACCEPT	accepting the proposal
ACCEPT-PART	accepting some part of the proposal
MAYBE	neither accepting nor rejecting the proposal
REJECT-PART	rejecting some part of the proposal
REJECT	rejecting the proposal
HOLD	putting off response, usually via subdialogue
ANSWER	answering a question
UNDERSTANDING	whether speaker understood previous
SIGNAL-NON-UNDER.	speaker didn't understand
SIGNAL-UNDER.	speaker did understand
ACK	demonstrated via continuer or assessment
REPEAT-REPHRASE	demonstrated via repetition or reformulation
COMPLETION	demonstrated via collaborative completion

Slides from Dan Jurafsky and Paul Martin

Backward looking (2006)

- Agreement:
  - Speaker is addressing a previous proposal, request, or claim, with the possibility of accepting or rejecting all or part of the proposal, request or claim; of withholding his attitude towards the proposal, request, or claim; or stating his attitude while being non-committal to the proposal., request, or claim.
- Understanding:
  - Utterances concerning the understanding between Speaker and Hearer, ranging from merely hearing the words to fully identifying intention.
- Answer:
  - Standard reaction of Speaker to an Info-request action by Hearer.
- Information-relation:
  - Tag which should capture how the content of this utterance relates to the content of its antecedent (still subject of further study).

A DAMSL Labeling



[assert]	C_1:	. . . I need to travel in May.
[info-req,ack]	A_1:	And, what day in May did you want to travel?
[assert, answer]	C_2:	OK uh I need to be there for a meeting that's from the 12th to the 15th.
[info-req,ack]	A_2:	And you're flying into what city?
[assert,answer]	C_3:	Seattle.
[info-req,ack]	A_3:	And what time would you like to leave Pittsburgh?
[check,hold]	C_4:	Uh hmm I don't think there's many options for non-stop.
[accept,ack]	A_4:	Right.
[assert]		There's three non-stops today.
[info-req]	C_5:	What are they?
[assert, open-option]	A_5:	The first one departs PGH at 10:00am arrives Seattle at 12:05 their time. The second flight departs PGH at 5:55pm, arrives Seattle at 8pm. And the last flight departs PGH at 8:15pm arrives Seattle at 10:28pm.
[accept,ack]	C_6:	OK I'll take the 5ish flight on the night before on the 11th.
[check,ack]	A_6:	On the 11th?
[assert,ack]		OK. Departing at 5:55pm arrives Seattle at 8pm, U.S. Air flight 115.
[ack]	C_7:	OK.

Slides from Dan Jurafsky and Paul Martin

Automatic Interpretation of Dialogue Acts

- How do we automatically identify dialogue acts?
- Given an utterance:
  - Decide whether it is a QUESTION, STATEMENT, SUGGEST, or ACK
- Recognizing illocutionary force will be crucial to building a dialogue agent
- Perhaps we can just look at the form of the utterance to decide?

Slides from Dan Jurafsky and Paul Martin

Can we just use the surface syntactic form?

- YES-NO-Q’s have auxiliary-before-subject syntax:
  - Will breakfast be served on USAir 1557?
- STATEMENTs have declarative syntax:
  - I don’t care about lunch
- COMMAND’s have imperative syntax:
  - Show me flights from Milwaukee to Orlando on Thursday night

Slides from Dan Jurafsky and Paul Martin

Surface form != speech act type

	Locutionary Force	Illocutionary Force
Can I have the rest of your sandwich?	Question	Request
I want the rest of your sandwich	Declarative	Request
Give me your sandwich!	Imperative	Request

Slides from Dan Jurafsky and Paul Martin

Dialogue act disambiguation is hard! Who's on First?

Abbott: Well, Costello, I'm going to New York with you. Bucky Harris the Yankee's manager gave me a job as coach for as long as you're on the team.

Costello: Look Abbott, if you're the coach, you must know all the players.

Abbott: I certainly do.

Costello: Well you know I've never met the guys. So you'll have to tell me their names, and then I'll know who's playing on the team.

Abbott: Oh, I'll tell you their names, but you know it seems to me they give these ball players now-a-days very peculiar names.

Costello: You mean funny names?

Abbott: Strange names, pet names...like Dizzy Dean...

Costello: His brother Daffy Abbott: Daffy Dean...

Costello: And their French cousin.

Abbott: French?

Costello: Goofe'

Abbott: Goofe' Dean. Well, let's see, we have on the bags, Who's on first, What's on second, I Don't Know is on third...

Costello: That's what I want to find out.

Abbott: I say Who's on first, What's on second, I Don't Know's on third.

Dialogue Act ambiguity

- Can you give me a list of the flights from Atlanta to Boston?
  - This looks like an INFO-REQUEST.
  - If so, the answer is:
    - YES.
  - But really it's a DIRECTIVE or REQUEST, a polite form of:
  - Please give me a list of the flights...
- What looks like a QUESTION can be a REQUEST

Dialogue System Evaluation

- Key point about SLP.
- Whenever we design a new algorithm or build a new application, need to evaluate it
- Two kinds of evaluation
  - **Extrinsic:** embedded in some external task
  - **Intrinsic:** some sort of more local evaluation.

- How to evaluate a dialogue system?
- What constitutes success or failure for a dialogue system?

Slides from Dan Jurafsky and Paul Martin

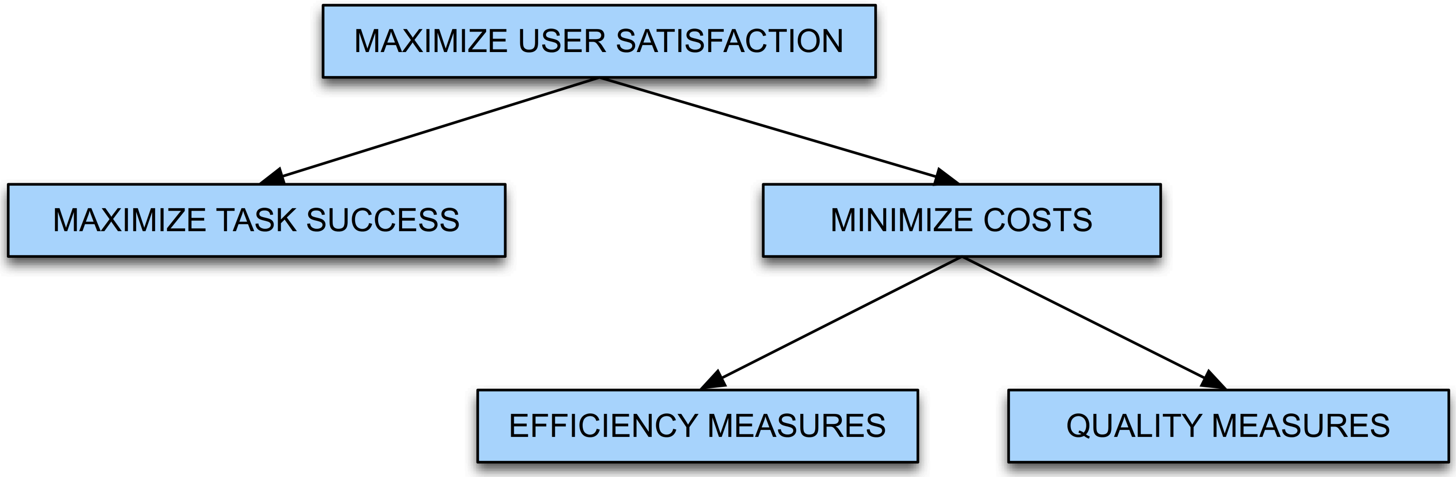
Dialogue System Evaluation

- It turns out we'll need an evaluation metric for two reasons
  - 1) the normal reason: we need a metric to help us compare different implementations
    - can't improve it if we don't know where it fails
    - Can't decide between two algorithms without a goodness metric
  - 2) a new reason: we will need a metric for "how good a dialogue went" as an input to reinforcement learning:
    - automatically improve our conversational agent performance via learning

Slides from Dan Jurafsky and Paul Martin

PARADISE evaluation

- Maximize Task Success
- Minimize Costs
  - Efficiency Measures
  - Quality Measures
- PARADISE (PARAdigm for Dialogue System Evaluation) (Walker *et al*. 2000)



Slides from Dan Jurafsky and Paul Martin

Task Success

- % of subtasks completed
- Correctness of each questions/answer/error msg
- Correctness of total solution
  - Attribute-Value matrix (AVM)
  - Kappa coefficient
- Users’ perception of whether task was completed

Slides from Dan Jurafsky and Paul Martin

Task Success

- **Task goals seen as Attribute-Value Matrix**

  ELVIS e-mail retrieval task ***(Walker et al '97)***

  "Find the time and place of your meeting with Kim."

  Kim c

Attribute	***Value***	***Selection Criterion***
Time	***10:30 a.m.***	
Place	***2D516***	

- **Task success can be defined by match between AVM values at end of task with "true" values for AVM**

Slides from Dan Jurafsky and Paul Martin

Slide from Julia Hirschberg

Efficiency Cost

- Polifroni et al. (1992), Danieli and Gerbino (1995) Hirschman and Pao (1993)

- Total elapsed time in seconds or turns
- Number of queries
- Turn correction ration: number of system or user turns used solely to correct errors, divided by total number of turns

Quality Cost

- # of times ASR system failed to return any sentence
- # of ASR rejection prompts
- # of times user had to barge-in
- # of time-out prompts
- Inappropriateness (verbose, ambiguous) of system’s questions, answers, error messages

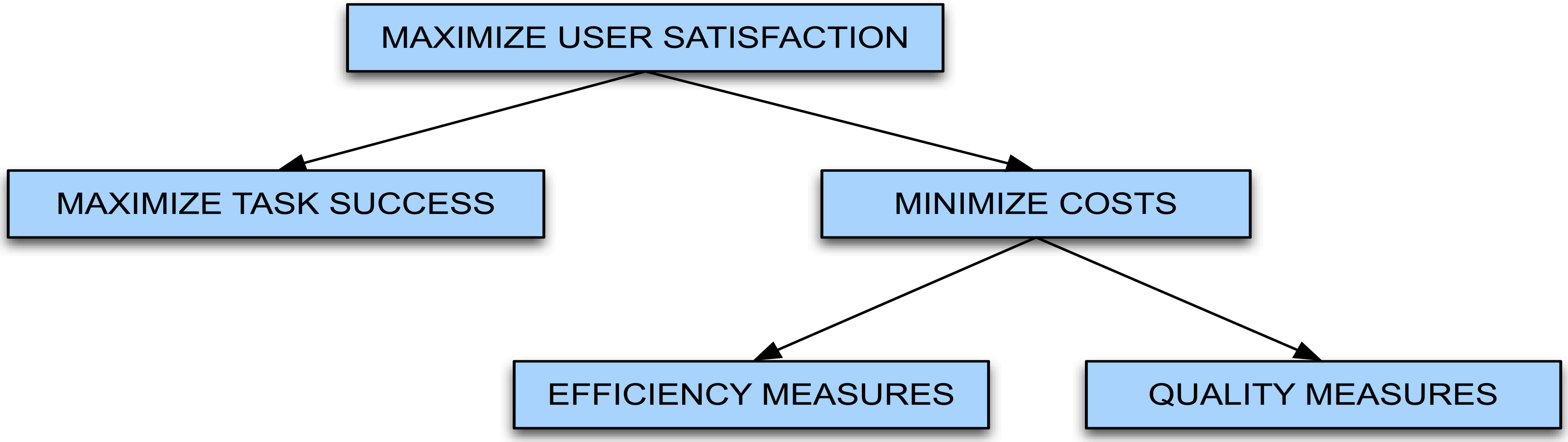
Slides from Dan Jurafsky and Paul Martin

Another key quality cost

- “Concept accuracy” or “Concept error rate”
- % of semantic concepts that the NLU component returns correctly
- I want to arrive in Austin at 5:00
  - DESTCITY: Boston
  - Time: 5:00
- Concept accuracy = 50%
- Average this across entire dialogue
- “How many of the sentences did the system understand correctly”

Slides from Dan Jurafsky and Paul Martin

PARADISE: Regress against user satisfaction



Regressing against user satisfaction

- Questionnaire to assign each dialogue a "user satisfaction rating": this is dependent measure
- Set of cost and success factors are independent measures
- Use regression to train weights for each factor

Slides from Dan Jurafsky and Paul Martin

Experimental Procedures

- Subjects given specified tasks
- Spoken dialogues recorded
- Cost factors, states, dialog acts automatically logged; ASR accuracy,barge-in hand-labeled
- Users specify task solution via web page
- Users complete User Satisfaction surveys
- Use multiple linear regression to model User Satisfaction as a function of Task Success and Costs; test for significant predictive factors

Slides from Dan Jurafsky and Paul Martin

Slide from Julia Hirschberg

User Satisfaction: Sum of Many Measures

- Was the system easy to understand? (TTS Performance)
- Did the system understand what you said? (ASR Performance)
- Was it easy to find the message/plane/train you wanted? (Task Ease)
- Was the pace of interaction with the system appropriate? (Interaction Pace)
- Did you know what you could say at each point of the dialog? (User Expertise)
- How often was the system sluggish and slow to reply to you? (System Response)
- Did the system work the way you expected it to in this conversation? (Expected Behavior)
- Do you think you'd use the system regularly in the future? (Future Use)

Slides from Dan Jurafsky and Paul Martin

Performance Functions from Three Systems

- ELVIS User Sat.= .21* COMP + .47 * MRS - .15 * ET
- TOOT User Sat.= .35* COMP + .45* MRS - .14*ET
- ANNIE User Sat.= .33*COMP + .25* MRS +.33* Help
  - COMP: User perception of task completion (task success)
  - MRS: Mean (concept) recognition accuracy (cost)
  - ET: Elapsed time (cost)
  - Help: Help requests (cost)

Slides from Dan Jurafsky and Paul Martin

Slide from Julia Hirschberg