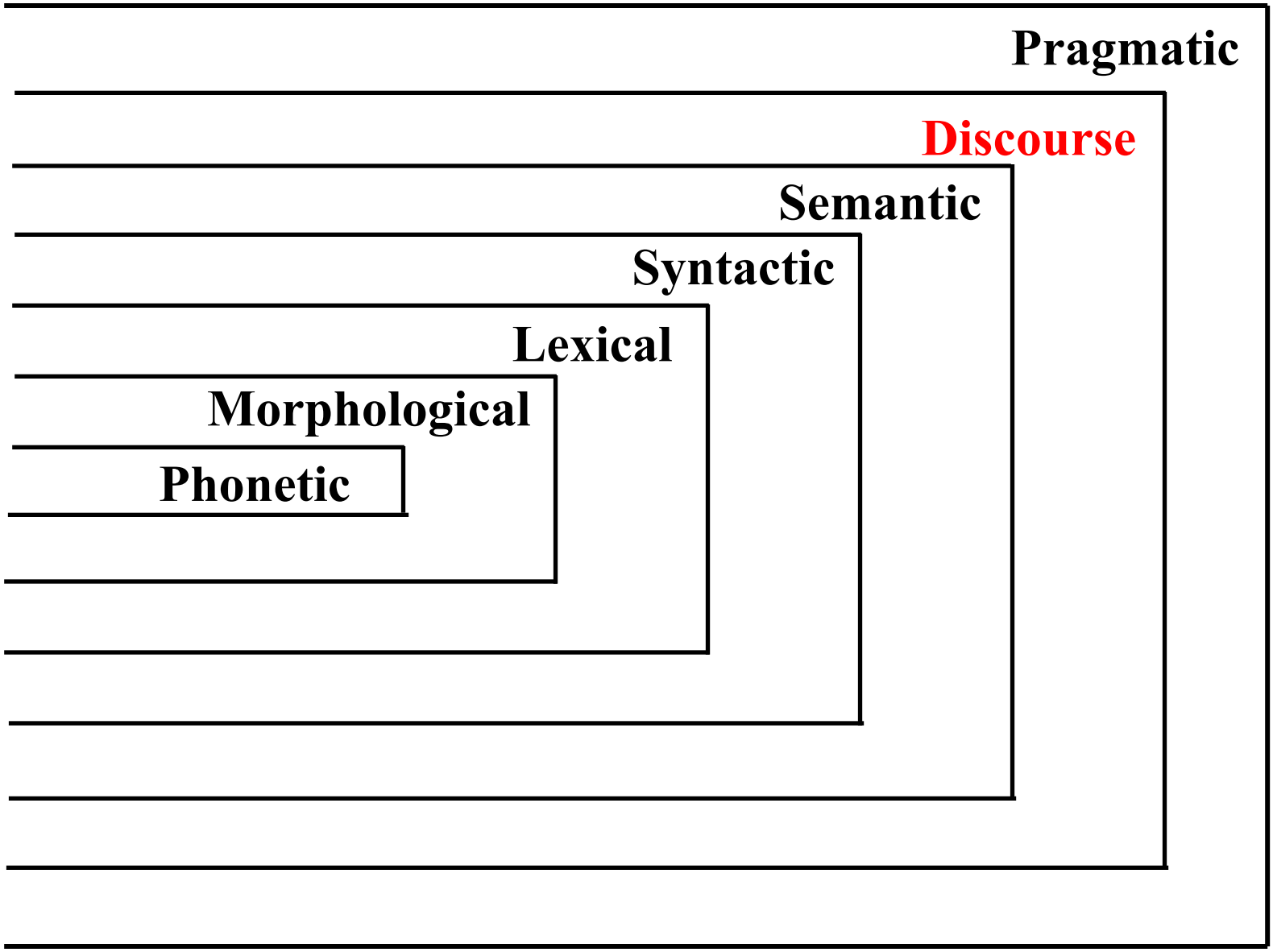
Introduction to Discourse Linguistics and Discourse Structure

Synchronic Model of Language



Discourse Linguistics

“ *No one is in a position to write a comprehensive account of discourse analysis. The subject is at once too vast, and too lacking in focus and consensus.* ” (Stubbs, Discourse Analysis)

Definitional Elements

- Study of texts (linguistic units) larger than a sentence.
- Text is more than a sequence of sentences to be considered one by one.
  - Rather, sentences of a text are elements whose significance resides in the contribution they make to the development of a larger whole.
- Each type of text has its own structure that can convey meaning to the reader.
- Some issues of discourse understanding are closely related to those in pragmatics which studies the real world dependencies of utterances.

Distinctions Between Text and Discourse

- In some contexts, e.g. in communication research, the word discourse means
  - interactive conversation
  - spoken
- And the word text means
  - non-interactive monologue
  - written
- But for (American) linguists, the word discourse can mean both of these things at the discourse level.

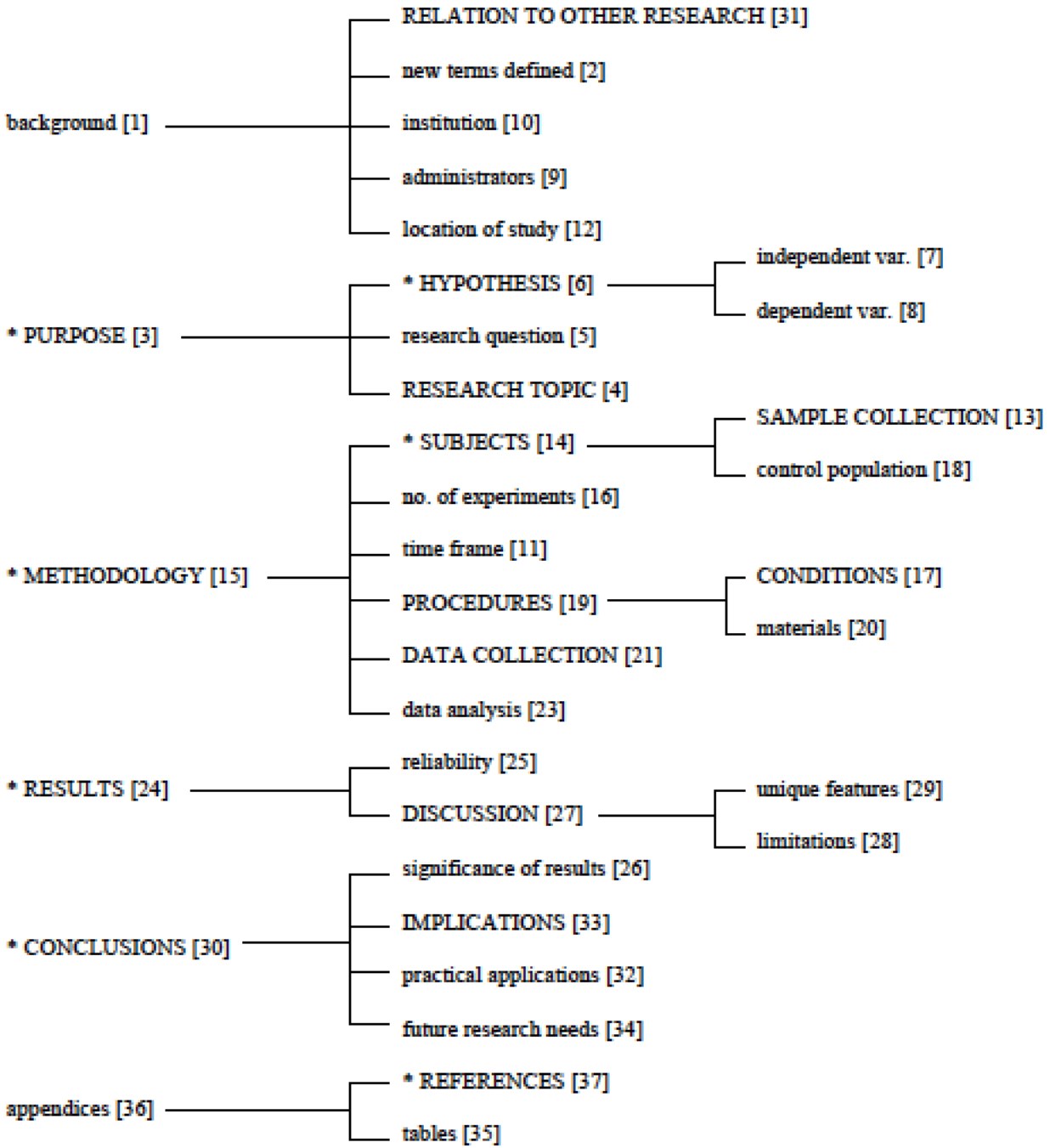
Scope of Discourse Analysis

- What does discourse analysis extract from text more than the explicit information discoverable by sentence-level syntax and semantics methodologies?
  - Structural organization of the text
  - Overall topic(s) of the text
  - Features which provide *cohesion* to the text
- What linguistic features of texts reveal this information to the analyst?

Discourse Structure

- Human discourse often exhibits structures that are intended to indicate common experiences and respond to them
  - For example, research abstracts are intended to inform readers in the same community as the authors and who are engaged in similar work
- Empirical study in dissertation by Liz Liddy identifies discourse structure of research abstracts
  - Hierarchical, componential text structure
  - Examples from Liddy discourse structure taken from Oddy, Robert N., "Discourse Level Analysis of Abstracts for Information Retrieval: A Probabilistic Approach", p. 22 - 23

Scheme of discourse elements for research abstracts



Research abstract with discourse elements marked

Empirical studies of Japanese work ethics have tended to focus on male workers while neglecting women. In addition, work values in both Japan and the United States appear to be changing. More information is needed on the work values of American and Japanese female workers.

BACKGROUND

A study was conducted to explore

the work ethics of Japanese women

RESEARCH TOPIC

PURPOSE

and to compare them to those of American women.

Subjects were 261 Japanese and 347 American employed women

SUBJECTS

who were tourists in Hawaii.

LOCATION

Subjects completed the Work Ethics questionnaire, an instrument designed to reflect the traditional values of both Japanese and American cultures. The questionnaire was translated into Japanese for Japanese subjects.

DATA COLLECTION

METHODOLOGY

T-tests used to test for significance of differences

DATA ANALYSIS

revealed that the Japanese and American women differed significantly on 27 of 37 work ethics. In comparison with American women, Japanese women were more prone to value group participation; to work in large rather than small companies; to value loyalty to employer and country; to desire more time for leisure and recreational activities; and to believe that suffering adds meaning to life and that money acquired easily is usually

RESULTS

Discourse Segmentation

- Documents are automatically separated into passages, sometimes called fragments, which are different discourse segments
  - Discourse segments can inform semantic interpretation of document
- Techniques to separate documents into passages include
  - Rule-based systems based on clue words and phrases
  - Probabilistic techniques to separate fragments and to identify discourse segments (Oddy)
  - Lexical cohesion to identify fragments (TextTiling)

TextTiling

- Uses lexical cohesion to identify segments, assuming that each segment exhibits "lexical cohesion" within the segment, but is not cohesive across different segments
- Algorithm
  - Identifies candidate segments
  - Computes lexical cohesion score in each segment –
    - Lexical cohesion score is the average semantic similarity of words within a segment
  - Identify boundaries by the difference of cohesion scores
  - NLTK has a text tiling algorithm available