

Paradigms, Corpora and Tools in Discourse
and Dialogue Research:
Corpus Creation as Understanding

Syun Tutiya
(Chiba University)

What I might want to say:

- Need for “paradigm”-independent dialog corpora →
 - Reusable
 - Retaggable(Reannotatable)
 - Extensible
- Requires an abstract notion of “corpus”
- The process of dialog corpus creation suggests a way of looking at the process of understanding dialog
- Inflation of “tools” is not harmful but productive
- Multiparty corpora are full of challenges

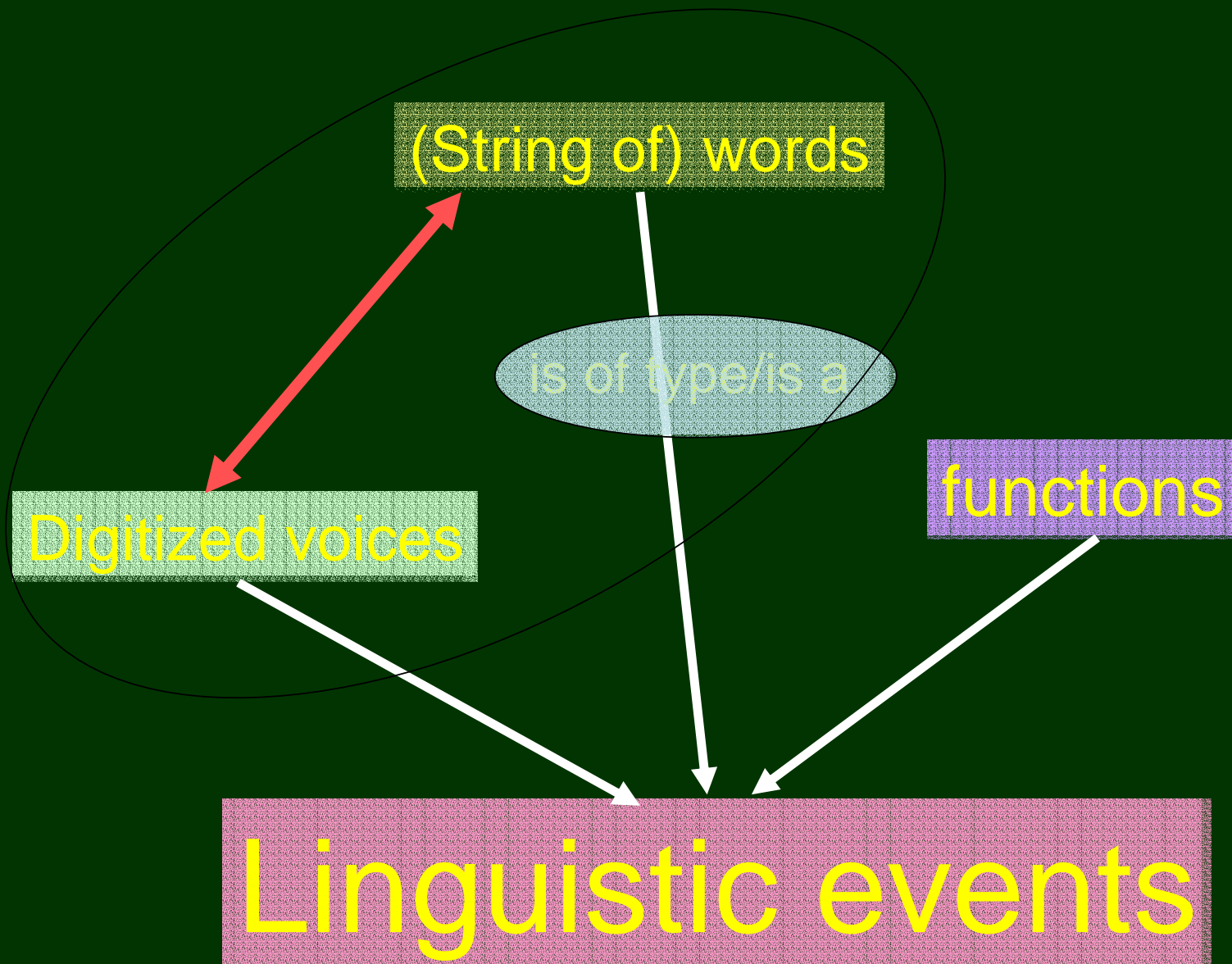
The purpose of corpora of all kinds

- Assume that any linguistic corpus is, in abstract terms, a triple, $\langle G, P, R \rangle$, where
 - G: Grounding information, i.e., time, space, participating agents, languages in use, moods and feelings, etc (Pragmatics)
 - P: Perspectives from which to look at the grounded information, like phonemes, prosodies, morphemes, words, phrasal structures, “meanings,” speech acts, intentions, discourse segments, etc (Semantics)
 - R: Relations between various perspectives. Grammar can be described in terms of morphemes and their orders, etc (Syntax)

Dialog corpora

- G: spatio-temporal (perhaps scattered) regions and participants
- P:
 - Digitized sounds
 - (“accompanying non-verbal behaviors”)
 - (phonemes)
 - Words and non-words
 - “Grammar”
 - Turn(?)
 - Speech acts
 - Discourse structures
 - Tasks(?), etc
- R: morphophonetics/Grammars/Speech acts/etc

dialog corpus as (contextual) dictionary



Corpus creation as understanding

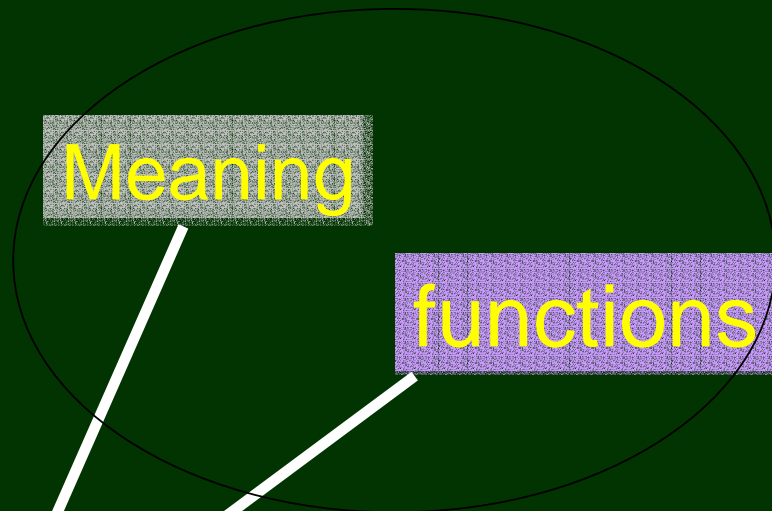
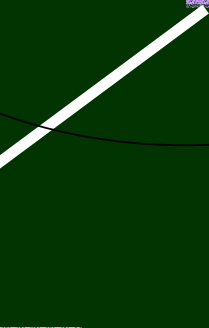
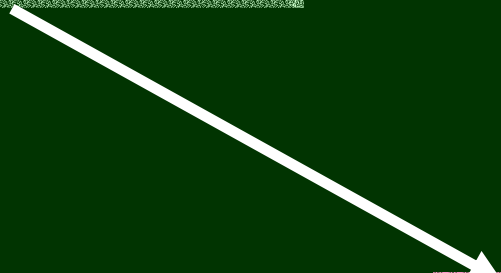
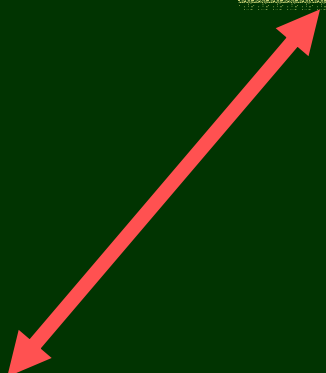
(String of) words

Meaning

functions

Digitized voices

events



Corpus as understanding

- Corpus is a representation, or a organized set of representations of “events in the world”
- Therefore, creation of corpus is understanding of events
- Therefore, creation of dialog corpus is dialog understanding
- From experiences in corpus creation, lessons are:
 - Accumulation of “tags” from different perspectives (case of speech act/intention tagging)
 - The complete picture comes out after corpus is created
 - No one way processing, but leveraged increase of “understanding” from different perspectives, levels, layers, ...
- A new paradigm?

From twoparty dialog corpus to multiparty dialog corpus

- Can TPDC be extensible compositionally to MPDC?
 - Multiparty dialogs as composition of twoparty dialogs organized by the structure of the tasks to be done by as many participants
- Or, is TPDC a degenerate instance of MPDC?
 - Parameters which are conspicuous in multiparty dialogs have been latent even in twoparty dialogs
- Decision to be made by the creators of corpora