COLING 2020 Tutorial

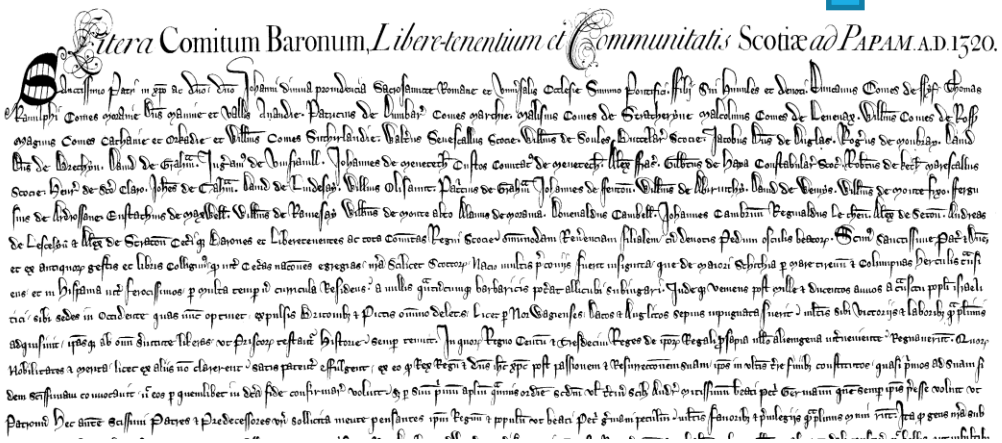
Cross-lingual Semantic Representation for NLP with UCCA:
A Bird's Eye View

Omri Abend

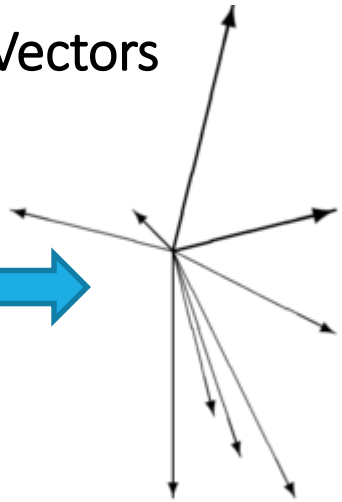
Semantic Analysis in NLP

Logical Forms

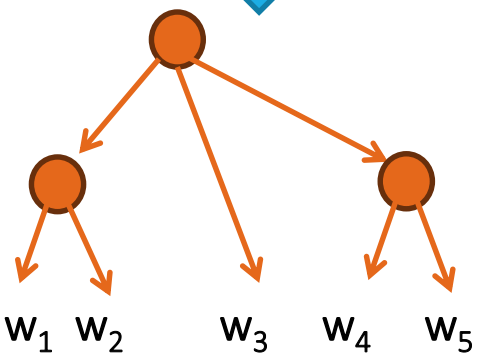
$\lambda x.p_1(a,b) \wedge p_2(c,x)$



Vectors



Trees/DAGs

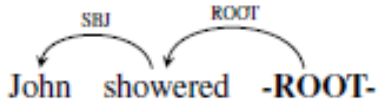


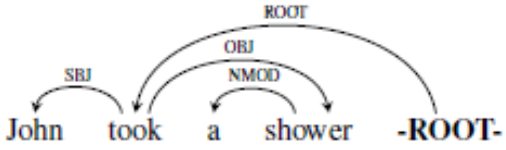
KBs

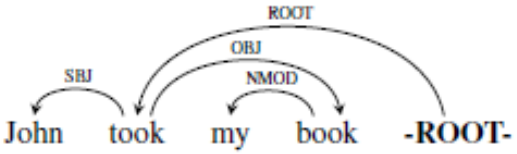
Symbolic Semantic Representation

- The focus of this tutorial: symbolic, sentence-level (or few sentences at most)
- By “semantic” NLPers refer to many different things
- Some of which:
  - Representation that supports inference
  - Representation that relates to the text to some extra-linguistic semantics (grounding)
  - The compositional structure of a sentence/text
  - An invariant of “meaning-preserving” variation (translation or paraphrase)

Semantic Structures: Stability to Paraphrasing







"John showered" ≈ "John took a shower" ≠ "John took my book"

Hebrew:
ג'ון התקלח
John showered-himself

Hebrew:
ג'ון לקח את
הספר שלי

Semantic Structures: Stability to Paraphrasing

Syntactic Schemes	Semantic Schemes
founding of the school	founding of the school
president of the United States	president of the United States
United States president	United States president

Why Symbolic Semantic Representation?

- Distributional methods (e.g., contextualized word embeddings) are very useful

- However:
  - They are difficult to interpret
  - It is difficult to read a compositional semantic account off them
  - A number of works have shown that even huge language models and other neural models can benefit from incorporating structure

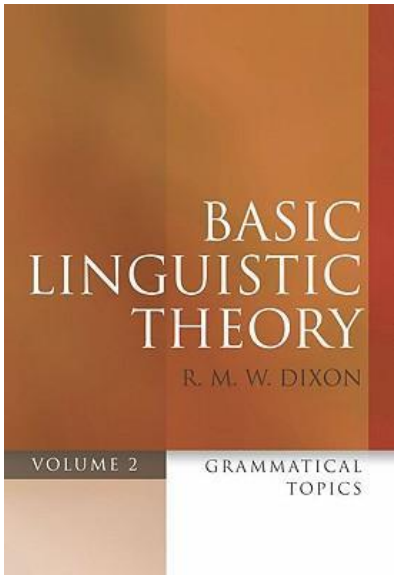
Why not just have Syntax?

- Syntactic structure is very useful, but
  - Syntactic schemes often under-specify, or are orthogonal to semantic distinctions
  - Syntax varies considerably across languages (translation divergences; e.g., Dorr, 1994)
- Accessibility to non-expert annotators
  - Syntactic annotation requires highly proficient annotators
  - Can semantic structure be more accessible?

UCCA: Design Principles

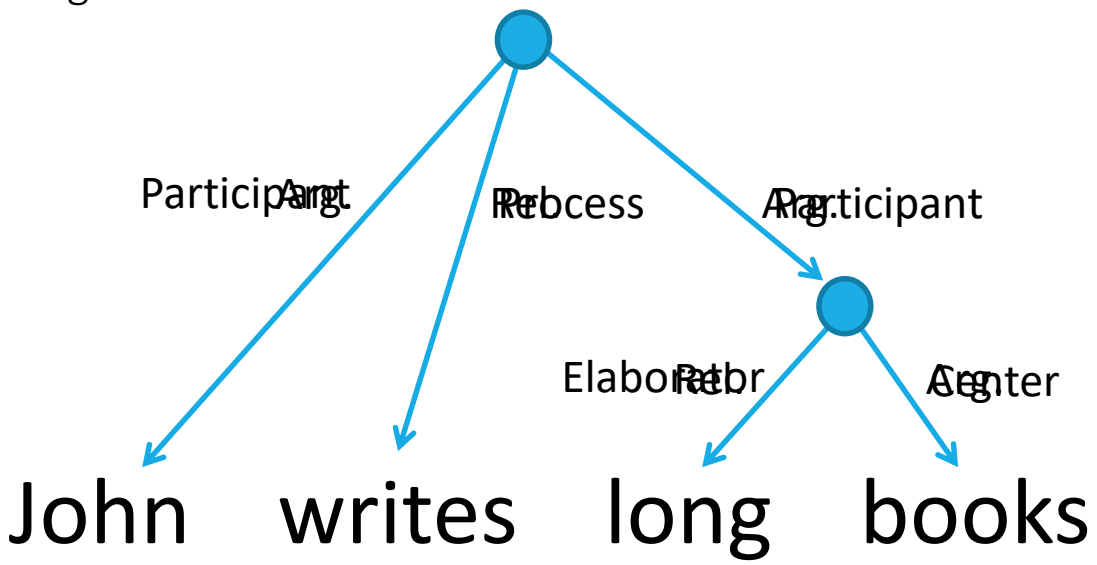
1. Abstract away from formal variation
2. Cross-linguistic applicability
3. Accessibility to non-expert annotators
4. Modularity

Theoretical foundations:
Mostly, *Basic Linguistic Theory*, a typological descriptive framework by *R.M.W. Dixon*



UCCA: Formalism

- Terminals
- Units
- Relations and arguments
- Categories
- Layers



UCCA's Foundational Layer

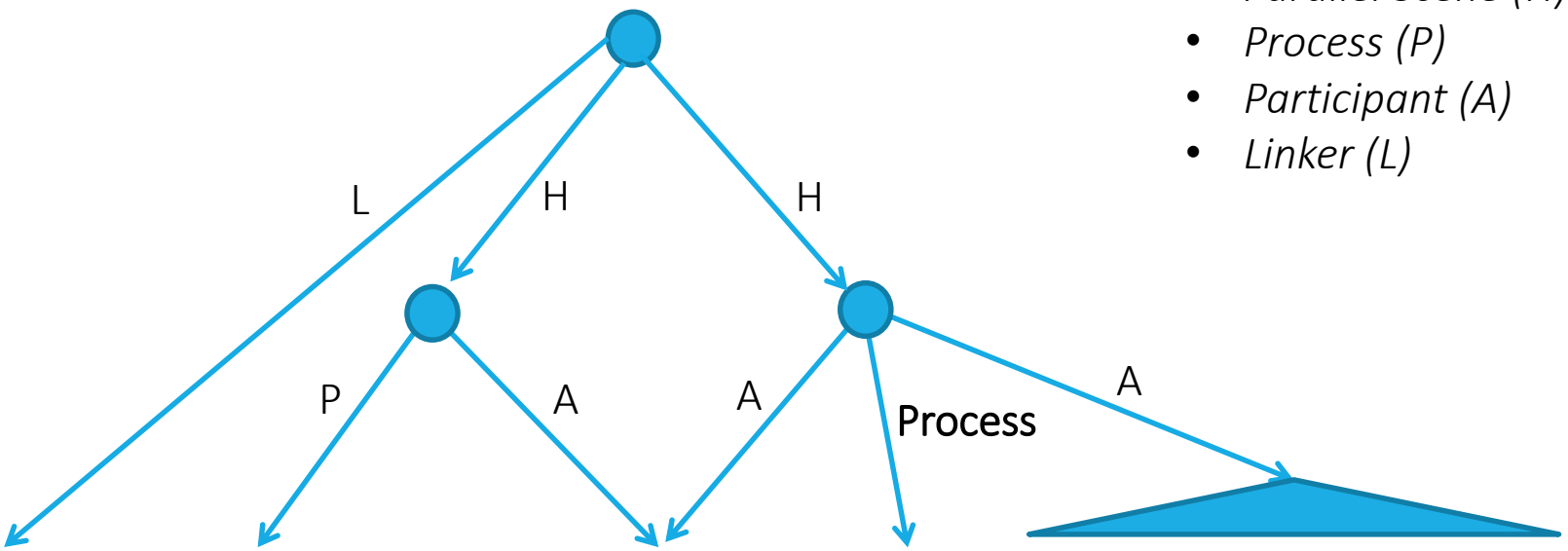
- **Focus:** semantic heads, predicate-argument relations and linkage between them
- Maximally coarse-grained (14 categories)
- Based on the semantic aspect of Basic Linguistic Theory's definition of a clause

UCCA's Foundational Layer: Scenes

"After graduation, Sorkin moved to New York City where he worked odd jobs including delivering telegrams, and driving a limousine."

UCCA's Foundational Layer: Scenes

- *Parallel Scene (H)*
- *Process (P)*
- *Participant (A)*
- *Linker (L)*

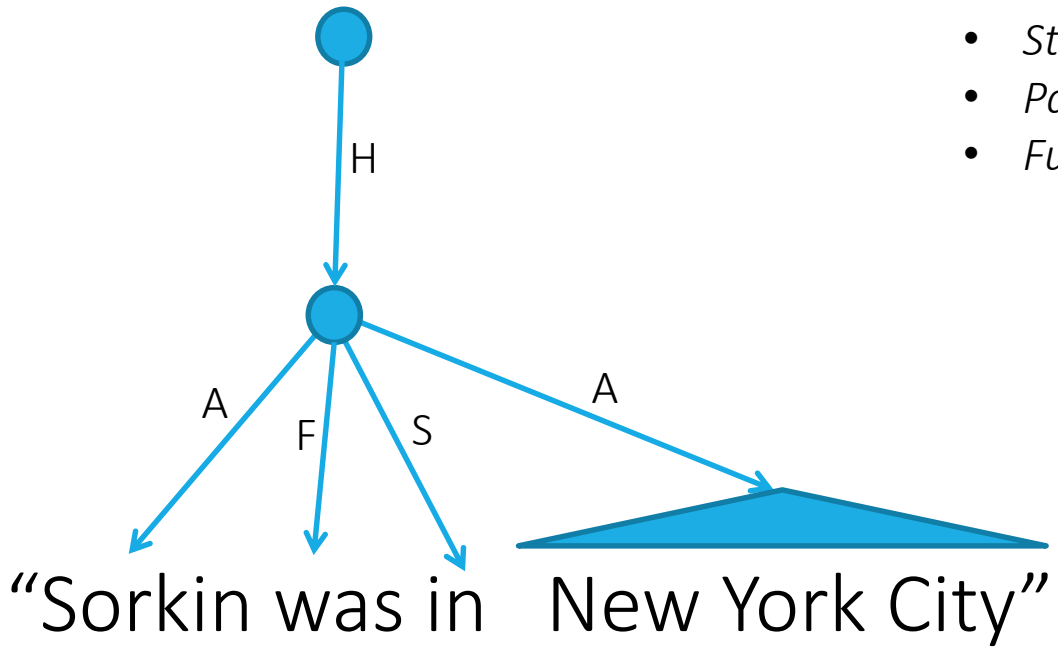


"After graduation, Sorkin moved to New York City"

- Nouns/adjectives/prepositions (etc.) can evoke scenes
- Participants need not be syntactic arguments

UCCA's Foundational Layer: Scenes

- *Parallel Scene (H)*
- *State (S)*
- *Participant (A)*
- *Function (F)*



- Nouns/adjectives/prepositions (etc.) can evoke scenes
- Participants need not be syntactic arguments

Secondary Verbs in UCCA

- English often uses verbs to express distinctions such as modality, aspect and causativity:
  - “**happen** to meet”, “**made** me laugh”, “**able** to sleep”
  - All of these are treated as standard verbs in UD and PTB

- A cross-linguistic perspective shows that these constructions vary considerably across languages:
  - “**happen** to meet” → German: “zufällig treffen” (lit. incidentally meet)
  - “made me laugh” → Hebrew: "הצחיק אותי" hicxik ʔoti (lit. made-laugh me)
  - “can sleep” → Japanese: **寝られる** nerareru (cf. **寝る** neru, trans. “sleep”)

Secondary Verbs in UCCA

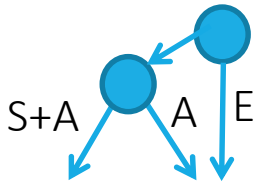
- UCCA annotates secondary verbs as *Adverbials (D)*
  - "can$_D$ go$_P$" and "[had to]$_D$ go$_P$"
  - "happen$_D$ to meet$_P$" → German: "zufällig$_D$ treffen$_P$" (lit. incidentally meet)
- A similar treatment is given to multi-word expressions that express these distinctions
  - "[**take a stab**]$_D$ at answering these"

  and to eventive nouns:
  - "the beginning$_D$ of the ceremony$_P$"

Sensitivity to Content, not Syntactic Categories

- Verbs can be adverbials:
  - "John **began**$_D$ swimming"
- Prepositions can be many things:
  1. Case markers: "Yossi lives **in**$_R$ Jerusalem"
  2. Linkers: "**After**$_L$ graduation, Sorkin moved to NYC"
  3. Scene-evokers: "The tree is **in**$_S$ the garden"

- Possessives can mark
  1. A State: "**your**$_{S+A}$ **car**$_A$"
  2. Part-whole relation: "**your**$_C$ **hands**$_C$"
  3. Participation: "**your**$_A$ **kick**$_P$"



Legend:
- *Process (P)*
- *State (S)*
- *Participant (A)*
- *Adverbial (D)*
- *Elaborator (E)*
- *Center (C)*

Inter-Scene Linkage

"After graduation, Sorkin moved to New York City where he worked odd jobs including delivering telegrams, and driving a limousine."

Scenes:

- "graduation$_P$ … Sorkin$_A$"
- "Sorkin$_A$ moved$_P$ [to New York City]$_A$"
- "he$_A$ worked$_P$ [odd jobs]$_A$"
- "he$_A$ … delivering$_P$ telegrams$_A$"
- "he$_A$ … driving$_P$ [a limousine]$_A$"

"and"

"after"

"where"

"including"

Coarse-grained, Refinable

- Two additional layers that refine the foundational layer:
  1. Semantic roles / Preposition supersenses (Schneider et al., 2018; Prange et al., 2019a):

     *Possession that is **not** scene-evoking:*
     - Kinship: "John's sister"
     - Part-Whole: "The car's windshield"

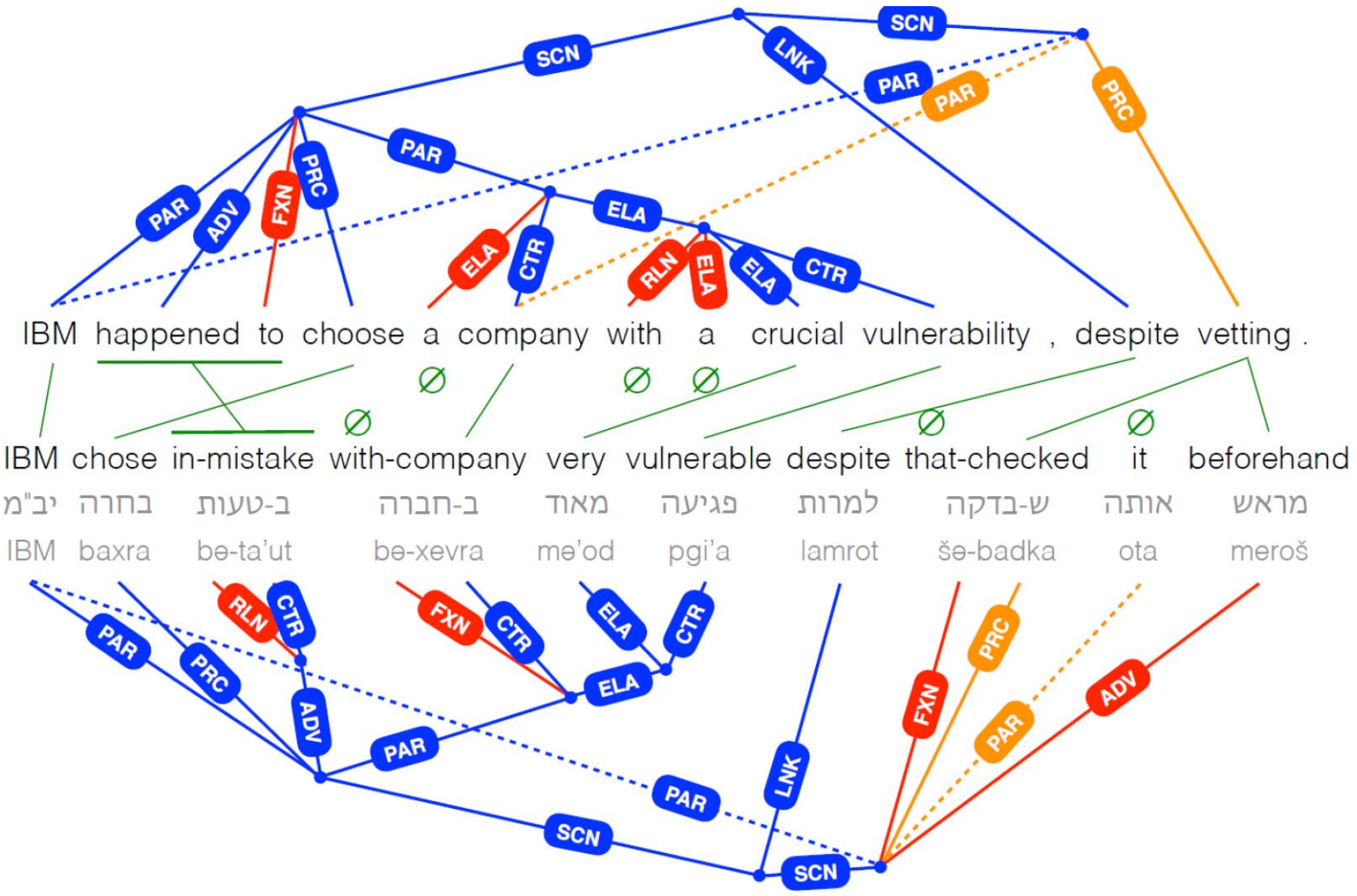
     Possession that is scene-evoking:
     - Agent: "John's kick saved the game"
     - Ownership: "John's computer"

  2. Coreference resolution (mentions are constrained by UCCA's foundational layer; Prange et al., 2019b)

Cross-linguistic Applicability

- UCCA aims to meet two goals:
  - **Portability**: the same set of categories and annotation guidelines can be applied across different languages
    - In practice: the set of categories and the bulk of the guidelines are shared, but per-language appendix is used to tackle the "long-tail"
  - **Stability** or structure preservation: a similar semantic structure is given to literal translations

Translation Divergences / Stability



UCCA Parsing in a Nutshell

- Three recent shared tasks:
  - SemEval 2019 shared task on UCCA parsing in English, French, German
  - CoNLL 2019 and 2020 shared tasks on Cross-Framework and Cross-Lingual Meaning Representation Parsing
- A number of parsers available, including
  - TUPA (Hershcovich et al., 2017): transition-based, multi-framework parser that served as baseline for two of the shared tasks
  - HLT@SUDA (Jiang et al., 2019): converting UCCA to constituency trees and a jointly-trained module for remote edges **(won the SemEval shared task)**
  - *Inter alia*
- Fairly mature technology for English, French, German. Hebrew and Russian in progress.

UCCA Applications in a Nutshell

- Evaluation of text2text systems:
  - Semantic measure for Grammatical Error Correction (Choshen and Abend, 2018)
  - Semantic measure for (structural) text simplification (Sulem et al., 2018)
  - Human evaluation guided by semantic structure for MT (Birch et al., 2016)
- Text simplification:
  - Text simplification using UCCA-based rules for preprocessing improves results (Sulem et al., 2018)
  - UCCA-guided simplification can also support MT in some settings (Sulem et al., 2020)
- Ongoing work on UCCA-based machine translation and relation extraction

Intermediate Summary

- Deep semantic analysis is increasingly important for NLP (despite advances in neural NLP)
- It can address a long-standing challenge: cross-linguistic applicability and stability

- I presented the *UCCA* approach:
  - Abstracts away from much syntactic variation
  - Demonstrated applicability to a number of languages
  - Corpora and parsers available for a number of languages
  - Already showing utility in evaluation of text2text systems and applications such as sentence simplification