Corpus Construction for Theoretical Analysis and NLP Applications: The Syntactic Corpus of English VP Ellipsis

Lotus Goldberg and Amber Stubbs

NASSLLI 2022, Day 3 – Wednesday June 22, 2022

Our course's website: https://sites.google.com/simmons.edu/vpe-corpus/nasslli2022

If you haven't yet, please visit our course website and fill out the 'First Day Survey'.
Thanks so much!

Plan for today: Phase 1 annotation process to detect instances of VPE

1. Pre-annotation work (detecting VPE instances)
2. Annotation environments
3. Phase 1 VPE annotation process in an nutshell, and a class exercise on Phase 1 annotation
4. Inter-annotator agreement
5. Discussing the data and results from the class Phase 1 annotation exercise

0. Some machine learning work with existing VP Ellipsis corpora

- Liu, Gonzalez, and Gillick (2016) "Exploring the steps of Verb Phrase Ellipsis"
    - Used Bos and Spenader (2011) + Nielsen (2005)

- Kenyon-Dean, Cheung, and Precup (2016) "Verb Phrase Ellipsis Resolution Using Discriminative and Margin-Infused Algorithms"
    - Used only Bos and Spenader (2011)

- Zhang et al (2019) "A neural network approach to verb phrase ellipsis resolution."
    - Used only Bos and Spenader

1. Decision on method for detecting VPE instances

- Slightly automated (highlighting triggers) and otherwise manual Phase 1 rather than more automated detection with regular expressions (such as used by Anand, Hardt, and McCloskey 2021 in creating the Santa Cruz Sluicing Data Set).

2. Setting up an annotation environment

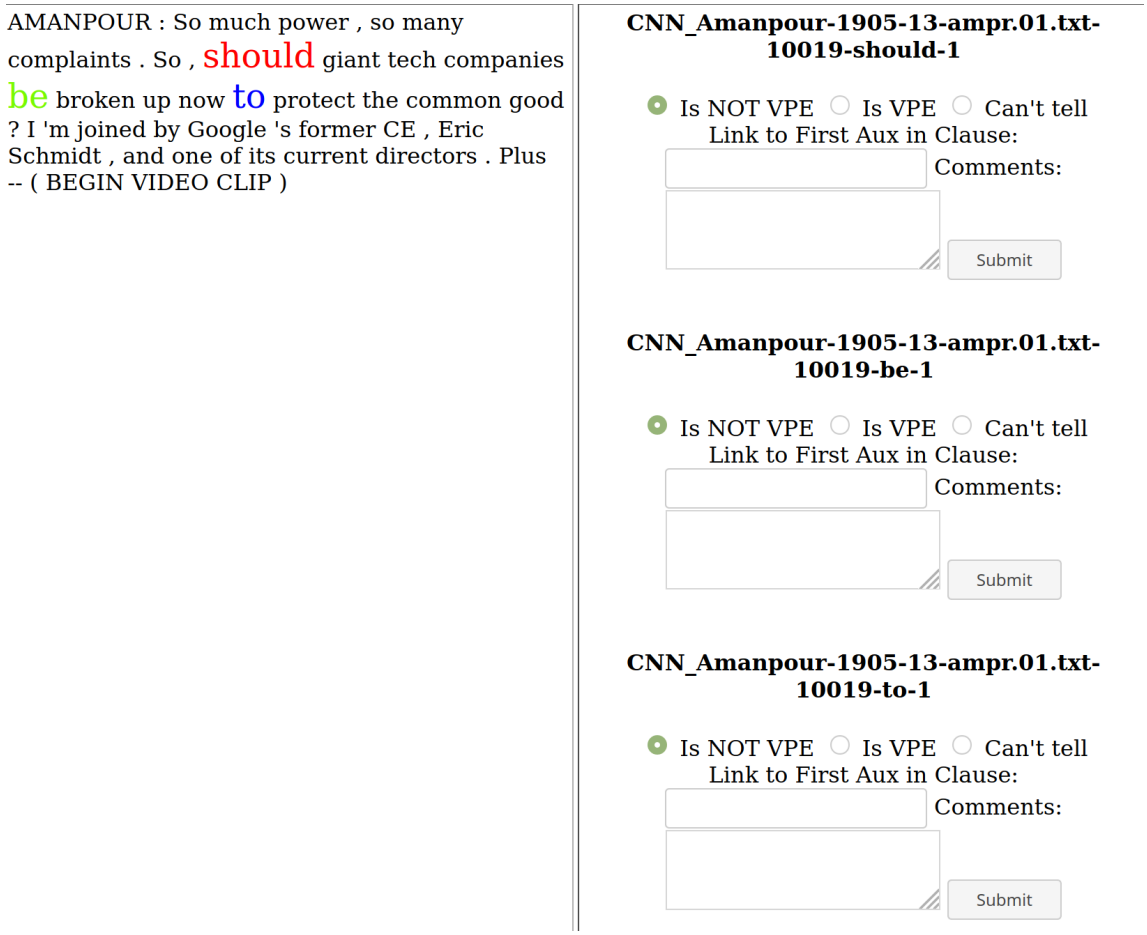
- Overview of annotation environments

    - Some commonly used annotation tools:
        - Apache UIMA: https://uima.apache.org/

        - brat: http://brat.nlplab.org/
        - GATE: https://gate.ac.uk/
        - Stanford Core NLP: https://stanfordnlp.github.io/CoreNLP/

    - A less commonly used annotation tool:
        - MAE (original flavor): https://github.com/amber-stubbs/mae-annotation
        - MAE (remixed): http://keighrim.github.io/mae-annotation/

- Specifics of our Phase 1 setup for identifying VPE



- Default assumption is "NOT VPE"
- Potential VPE triggers are highlighted by class
- Annotators submit a judgement for each trigger, then submit the completed document

3. Phase 1 annotation, in a nutshell

- Annotators load a website on which they click each file, showing them all text in the file

- For each sentence, all VP Ellipsis 'triggers' appear highlighted in different colors, depending on the trigger type:

- Modal Auxs (*could*, *can*, *will*, *would*, *should*, etc.)
- Forms of *have* (whether the Perfect Aux or main V (=copula))
- Forms of *be* (whether the Progressive Aux, the Passive Aux, or the main V)
- Forms of *do* (whether the pleonastic *do* of *do*-support, or the main V)
- *to* (so that we include infinitival *to*, but also including the P *to*)
- *not*, included so that we capture the two types of example noted by Lobeck (1995) to have no possible inflectional licenser for VPE other than clausal negation:
  - Cases where an Aux or copula is cliticized onto the preceding word:

    (1) Michael is arriving after dinner, but Harriet's *not* ~~arriving after dinner~~.

    cf. (2) *Michael is late, and Harriet's ~~late~~ too.

    (3) Michael is late, and Harriet is ~~late~~ too.

  - Small clauses, where there's no Aux, copula, or infinitval *to* present to serve as licensor:

    (4) We consider [$_{\text{Small Clause}}$ Henry a friend], but Alex not ~~a friend~~.

- Annotators evaluate each trigger with respect to whether it is part of an immediate clause containing an instance of VP Ellipsis or not, and:
  - Mark (a) Is VPE, (b) Isn't VPE, or (b) Can't tell
  - Note cases where there is more than one trigger for an instance of VPE
  - Can enter a comment, iff they find anything especially puzzling or confusing

- **Core criteria for marking a clause's trigger as being in an instance of VPE:**
  - At least one stranded inflectional element is present
  - A syntactically-defined VP is missing
    - As noted earlier this week, this can include copula clauses, in which what has elided will appear to be just a Noun Phrase, AP, or PP, with the copula (=main V *be*) left **overt**:

      (5) Cory is a student of linguistics / incredibly brilliant / over by the lake, and Sam **<u>is</u>**$_i$ [$_{\text{VP}}$ t$_i$ ~~a student of linguistics / incredibly brilliant / over by the lake~~] too.

      cf. (6) *Cory is a student of linguistics / incredibly brilliant / over by the lake, and Sam **<u>does</u>** [$_{\text{VP}}$ ~~be a student of linguistics / incredibly brilliant / over by the lake~~] too.

4. Class participation: VPE annotation

- Link to annotation form: https://forms.gle/h1rwkh6HG3aoX7TK9

— 5 Minute Break —

5. Calculating inter-annotator agreement (IAA)

- Key terms:
  - Precision, recall, and F-measure
    - Precision= # of true positives / (# of true positives + # of false positives)
    - Recall = # of true positives / (# of true positives + number of false negatives)
    - F-measure (harmonic mean of precision and recall)

$$F = 2 \times \left(\frac{p \times r}{p + r} \right)$$

  - Cohen's kappa
    - measures the agreement between two annotators, accounting for the possibility of chance agreement
  - Fleiss kappa
    - Measures the agreement between more than two annotators, accounting for the possibility of chance agreement

- Why is IAA important?

- IAA discussion for Phase 1 in our corpus
  - Limited amount of VPE compared to number of trigger words
    - Large number of rue negatives artificially increases F1 measure: .994
  - Focusing on recall

6. Syntactic issues regarding what should(n't) be considered VP Ellipsis

- Class discussion of the examples just annotated: is each an instance of VP Ellipsis or not, and why or why not?

- Larger points about what to include vs. exclude. (See the current version of the VPE Phase 1 Annotation Addendum/FAQ, which you can access in the Google Drive folder of additional materials from the course website, for good examples here.)

  - not *do so* (but YES to *so* + one or more Auxs/Copula)

  - not gaps that aren't VPs (e.g. relative gaps, wh-question gaps, missing material within gapping constructions)

- no clauses that lack a trigger
- no clauses in which a trigger occurs without a plausible elided VP, including in idioms such as *if I may*, *as it were*, *if you will*.

7. Overview of the class Phase 1 annotation results

8. Today's take-home question:

- What would your annotation scheme look like?

References

Anand, Pranav, Daniel Hardt & James McCloskey. 2021. The Santa Cruz Sluicing Data Set. *Language* 97(1), e68–e88. doi:10.1353/lan.2021.0009.

Bos, Johan & Jennifer Spenader. 2011. An Annotated Corpus for the Analysis of VP Ellipsis. *Language Resources and Evaluation* 45(4). 463–494.

Kenyon-Dean, Kian, Jackie Chi Kit Cheung & Doina Precup. 2016. Verb Phrase Ellipsis Resolution Using Discriminative and Margin-infused Algorithms. In *EMNLP 2016 - Conference on E Methods in Natural Language Processing, Proceedings*, 1734–1743. https://doi.org/10.18653/v1/d16-1179.

Liu, Zhengzhong, Edgar Gonzàlez Pellicer & Daniel Gillick. 2016. Exploring the Steps of Verb Phrase Ellipsis. In *Proceedings of the Workshop on Coreference Resolution beyond OntoNotes (CORBON 2016), co-located with NAACL 2016*, 32–40. https://doi.org/10.18653/v1/w16-0705.

Nielsen, Leif Arda. 2005a. A Corpus-Based Study of Verb Phrase Ellipsis Identification and Resolution. Dissertation.

Zhang, Wei Nan, Yue Zhang, Yuanxing Liu, Donglin Di & Ting Liu. 2019. A neural network approach to verb phrase ellipsis resolution. *33rd AAAI Conference on Artificial Intelligence, AAAI 2019, 31st Innovative Applications of Artificial Intelligence Conference, IAAI 2019 and the 9th AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019* 7468–7475. https://doi.org/10.1609/aaai.v33i01.33017468.