Butterfly Syndrome Corpus README

Spencer Thomas

November 27, 2022

The following is a fictitious corpus made to be situated in the universe of the podcast *Ars Paradoxica*. All of the persons, entities, and events introduced past this box are purely fictional.
With that said, this corpus was inspired by GRAEC (Kasselimis et al. 2020) and the Fromkin Speech Error Database. All of the data were obtained by hand-picking lines of dialogue from the transcripts on `https://arsparadoxica.com`.

0 Abstract

Butterfly Syndrome (BS) is one of the most common and most dangerous occupational hazards in the United States Office of Developed Anomalous Resources (ODAR). Its most identifiable symptom is tense-specific aphasia. We have compiled examples of the resultant speech errors into the Butterfly Syndrome Speech Error Dataset (BSSED) in the hopes that the data therein will be able to inform clinicians identify and researchers study BS.

1 Data

In 1955, Morales and Sharma (1955) began broadcasting a series of audio recordings that had been disseminated from the Blackroom from its anchor point in 1943. Recall that BS is a common hazard at ODAR; since these recordings had all passed through ODAR's communications nexus, then, it should not be surprising that many of them include speech from individuals with BS.

It is understood that the Blackroom has shared with the world every audio recording that had crossed its desk. Morales and Sharma (1955) have shared a small portion of this data, about 18 hours' worth of speech. We hope that these actors will publish more recordings in the future, but in the meantime, their existing broadcasts have been transcribed by the *Ars Paradoxica* team.

The utterances used in this corpus were hand-selected from these transcripts. Given a list of individuals known to have BS, annotators located these individuals' utterances that displayed the disfluencies now known as signs of the disorder.

All of the data in this corpus are freely available. Anyone who listens to the broadcasts in question or reads their transcripts has access to the knowledge contained herein. That includes patients' names, severity of illness, treatment, and other sensitive details. Even so, information that can be used to identify subjects has been omitted.

Data has been collated into XML files: one per subject.

2 Annotation

In our pilot work, we identified three types of speech errors: tense-aspect-mood (TAM), lexical, and omission.

2.1 Tense-Aspect-Mood

Erroneous utterances are listed by `instance`. These are then sorted by type, as described below. An `instance` is structured as follows:

- Instance
  - "episode": the "episode" of Morales and Sharma's broadcast that the instance appears in
  - Utterance: the sentence in question.
    - Precedent: if relevant, the preceding sentence or phrase.
  - Actual: the incorrect phrase uttered. There may be more than one of these.
    - "explicit": whether the full utterance is present
    - "tense": past, present, future, or unknown
    - "aspect": unmarked, perfect, progressive, unknown, or some combination of the above
    - "voice": active, passive, or unknown
    - "target span": character span of the target phrase within the utterance
    - "aux": the auxiliary verbs, modals, negatives, and infinitival "to"s used in the target phrase
  - Correct: the correct phrase. This may not actually be uttered.
    - All features are the same as defined for *Actual* above.

```
<instance episode="06">
    <utterance>I will see... I've seen that, clear as day.
        <precedent>What you are making has limitless potential to change the world.</precedent>
    </utterance>
    <actual explicit="true" tense="future"    aspect="unmarked" voice="active" targetspan="0~9"
        aux="be">I will see</actual>
    <correct explicit="true" tense="present" aspect="perfect" voice="active" targetspan="14~23"
        aux="have">I've seen</correct>
</instance>
```

2.2 Lexical and Omission

While most examples consisted of tense-aspect-mood errors, a large minority involved other lexical mistakes. In these examples, subjects replaced words with less fitting ones (*lexical* errors) or omitted them entirely (*omission* errors). While *omission* errors are probably a subtype of *lexical* errors, there is no neat annotation schema that describes both adequately.

2.2.1 Lexical

A lexical `instance` is structured as follows:

- Instance
  - "episode": the "episode" of Morales and Sharma's broadcast that the instance appears in
  - Utterance: the sentence in question.
    - Precedent: if relevant, the preceding sentence or phrase.
  - Actual: the incorrect phrase uttered.
    - "explicit": whether the full phrase is present.
    - "targetspan": the character span of the target phrase.
  - Correct: the correct phrase. This is usually not uttered.
    - "explicit": whether the full phrase is present. As noted directly above, this is usually false.

```
<instance episode="06">
    <utterance>But only where you find the right answer.
        <precedent>We've talked about this.</precedent>
     </utterance>
    <actual explicit="true" targetspan="9~13">where</actual>
    <correct explicit="false">when</correct>
</instance>
```

2.2.2 Omission

An omission instance is structured as follows:

- Instance
  - "episode": the "episode" of Morales and Sharma's broadcast that the instance appears in
  - Utterance: the sentence in question.
    - Precedent: if relevant, the preceding sentence or phrase.
  - Actual: the incorrect phrase uttered.
    - "targetspan": the space where the omitted word should have been uttered. If the gap is at the beginning of a sentence, this is 0.
    - "missing": the constituents that are omitted. This is formatted as some combination of S (subject), V (verb), O (object), or PP (prepositional phrase). If more than one constituent is missing, they are separated with a space so that they may easily be separated in preprocessing. Other constituents have not been observed to be omitted thus far.
  - Correct: an approximately correct phrase, chosen by the annotator.

```
<instance episode="06">
    <utterance>Not acceptable.</utterance>
    <actual explicit="true" targetspan="0" missing="S V"></actual>
    <correct explicit="false">That is<correct>
</instance>
```

3 Findings

As of this dataset's most recent update, this dataset only contains examples from one patient. This patient, however, can demonstrate a good deal about BS's effect on speech. We have a good number of recordings of this subject spanning from their introduction to time travel in 1943, through the worsening of their symptoms up to 1946, and even some after very effective CAGE therapy nearly eliminated their condition. The disordered speech samples in this corpus were all uttered before this treatment.

This subject has exhibited more aphasic speech errors than many others with BS, and certainly more than the other subjects we have access to. We have access to 24 such instances. Unsurprisingly, this subject's TAM errors became more severe as their condition degraded. Toward the culmination of their symptoms, they would utter multiple incorrect auxiliaries multiple times before every verb.

We also found that most of this subject's lexical errors involved time-related words. Within one recording midway through the development of their condition, this subject made the following replacements:

Uttered	Correct
span	room
thinking	thoughts
where	when
next to	until

All of these except for *thinking*/*thoughts* are clearly time-related.

Most intriguingly, we found that in most of this subject's TAM errors (10 of 17), the subject was *I* or *we*. We have yet to perform a full analysis of this subject's speech, but to the naked eye, it seems that most of their statements have other named entities as subjects. BS is caused by a break in one's perception of causality, particularly in relation to *oneself*. It makes sense, then, that a patient would have more difficulty with TAM when talking about oneself.

4 Future work

Our next plan of action is to collate data like the above for other subjects in Morales and Sharma's broadcasts. We would also like to create a corpus of these patients' non-disordered speech, so that we can directly compare it with their BS-affected speech. This will be a larger undertaking, but not by much; we can still obtain this data rather simply from Morales and Sharma's broadcasts.

Given the small amount of data we have available, it is unlikely that this dataset will ever be large enough to train machine learning models. However, we believe it is worthwhile to compile this data nonetheless. Even at this elementary stage, we have observed multiple heretofore unnoticed patterns in our subject's speech. We hope that these observations, and ones gleaned later on in our research, will help clinicians identify, diagnose, and treat BS.

5 References

Kasselimis, Dimitrios, Maria Varkanitsa, Georgia Angelopoulou, Ioannis Evdokimidis, Dionysis Goutsos, and Constantin Potagas. "Word error analysis in aphasia: introducing the Greek aphasia error corpus (GRAEC)." Frontiers in psychology (2020): 1577.