PROOF OF CONCEPT FOR SPOKEN NATURAL LANGUAGE DESCRIPTIONS OF PHENOTYPES FOR ASSOCIATION GENETICS APPLICATIONS

ABSTRACT

Imagine walking through a field and saying aloud what you see, then using that audio file to conduct an association study. Advancements in Natural Language Processing (NLP) have allowed for processing large volumes of descriptive data. To make these advancements applicable for biologists interested in better understanding phenotypes and traits, we are developing methods to collect and process in-field descriptions of maize using recordings of spoken phenotype descriptions. We planted the Wisconsin Diversity panel in Boone, Iowa (summer of 2021). Nine undergraduate student workers recorded spoken descriptions of the Wisconsin Diversity panel lines in the field. We instructed the students to describe certain plant parts and other miscellaneous attributes using their own words. For comparative purposes, we also collected numerically scored phenotypes using traditional data collection techniques and images of each line. Our pipeline for processing the hundreds of hours of spoken descriptive data starts with the Amazon Web Services (AWS) cost-effective transcription service Transcribe. We can compute on the text descriptions of plants produced by Transcribe using NLP tools, which enables us to generate the input for tools that are commonly used by the maize research community to perform Genome-Wide Association Studies (GWAS). The results of the association studies completed using spoken data are then compared to published associations. Generating spoken descriptive datasets and protocols for demonstrating biologically relevant approaches for these data are anticipated to enable the maize research community to use innovations in language processing for in-field phenotyping methods.

METHODS

Follow the yellow arrows to review the methods we are using to process spoken recordings to perform GWAS.

STEP 1: SPEECH-TO-TEXT

Transcription Processing Workflow

Record Speech for Plants of Each Genotype in the Field → AWS Transcribe → Computable Speech-Derived Text Observations

Observation Example

delta 1181 these corn plants are short … no visible brace roots → `delta 1181 these corn plants are short … no visible brace roots`

zulu 1360 overall appearance is kind of bushy … japanese beetles on ears as well → `zulu 1360 overall appearance is kind of bushy … japanese beetles on ears as well`

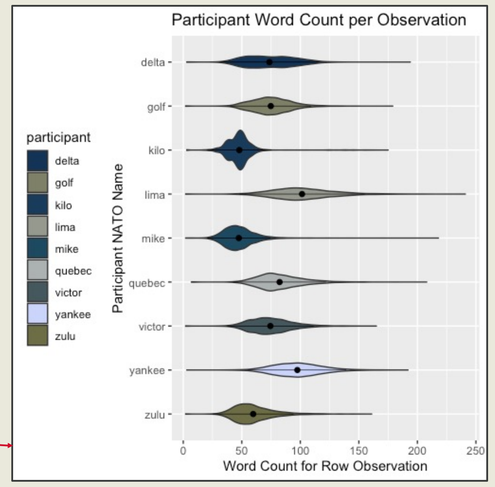
AWS Processing & Observation Summaries

Summaries of Participant Audio Data & Observations

Participant	Estimated Recording Duration	Estimated AWS Processing Time	Estimated AWS Cost	Total Rows Observed	Mean Words Per Row
Delta	~ 32:39:08	~ 10:53:23	~ $ 47.02	4,326	~ 73.61
Golf	~ 35:40:42	~ 07:49:46	~ $ 51.38	4,317	~ 74.79
Kilo	~ 27:54:24	~ 07:48:02	~ $ 40.20	4,321	~ 47.90
Lima	~ 47:09:47	~ 11:52:33	~ $ 67.94	4,322	~ 101.48
Mike	~ 29:08:29	~ 08:37:56	~ $ 41.98	4,302	~ 47.46
Quebec	~ 39:14:18	~ 09:13:55	~ $ 56.52	4,308	~ 82.28
Victor	~ 35:56:02	~ 08:43:53	~ $ 52.06	4,329	~ 74.43
Yankee	~ 44:10:55	~ 08:12:29	~ $ 64.47	4,308	~ 97.45
Zulu	~ 52:37:44	~ 09:47:23	~ $ 75.81	4,314	~ 59.73
Mean	~ 38:16:49	~ 09:13:16	~ $ 55.26	~ 4,316	~ 73.24
Total	~ 344:31:30	~ 82:59:20	~ $ 497.38	38,847	-

Curated Observation Dataframe

Participant	Observation Date	Row Number	Taxa	Observation
delta	07_02_21	1181	ZS1791	these corn plants are short … no visible brace roots
…	…	…	…	…
zulu	07_22_21	1360	LH38	overall appearance is kind of bushy … japanese beetles on ears as well



Distribution of Participant Word Count for Each Observation

Distribution of word count for nine student participants for all observed word counts. The sphere within the violin plot indicates the mean word per row; horizontal lines indicate the mean times the standard deviation.

STEP 2: TEXT-TO-SIMILARITY SCORE

HYPOTHETICAL EXAMPLE OF METHOD 1:

Query String: "tall plant" → **spaCy Pre-Processing:** `"tall plant"`

Observation Strings:
- "these corn plants are short … no visible brace roots" → `"these corn plant be short ... no visible brace root"`
- …
- "overall appearance is kind of bushy … japanese beetles on ears as well" → `"overall appearance be kind of bushy ... japanese beetle on ear as well"`

`similarity()` → Calculate similarity score between the Query & the Observation using spaCy

Similarity Score for each Observation: 0.57541453 … 0.62569073

Mean Similarity Score for Each Taxa

Taxa	Mean Similarity Score
ZS1791	0.58310524
…	…
LH38	0.57814002

The text-to-similarity score step (STEP 2) uses the **spaCy** Python library for:

Pre-processing: removing punctuation, removing stop words, & word lemmatization ("base form" or "dictionary form")

Method 1: Computing cosine similarity vectors using the spaCy similarity function

Method 2: Performing noun chunking to determine noun phrases using spaCy for its noun chunking iterator

HYPOTHETICAL EXAMPLE OF METHOD 2:

Observation Strings:
- "these corn plants are short … no visible brace roots" → `"these corn plant be short ... no visible brace root"`
- …
- "overall appearance is kind of bushy … japanese beetles on ears as well" → `"overall appearance be kind of bushy ... japanese beetle on ear as well"`

Determine noun chunks using spaCy

`noun_chunks`

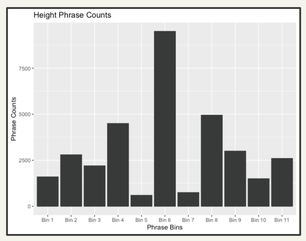
```
small stature,
small plant,
...
average height,
average length,
...
tall plant,
super tall height
```

short height → tall height

Height Phrase Determination

Bin Number	Phrase(s)
1	"very small" & "very little" & "extremely small"
2	"really small"
3	"kind of small" & "a bit short"
4	"smaller than average" & "medium small"
5	"not tall"
6	"relatively average" & "average length" & "average height" & "average length"
7	"not short"
8	"taller than average" & "medium tall"
9	"somewhat tall" & "kind of tall" & "relatively tall" & "quite tall"
10	"really tall"
11	"super tall" & "extremely tall"

Score Height Phrase



short height → tall height

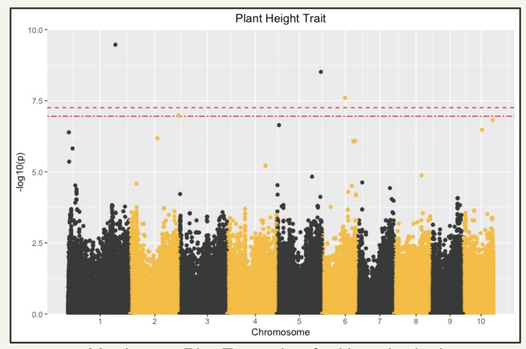
0 → 1

Score Scale for Height Phrase Bins

Mean Score for Each Taxa

Taxa	Mean Score
ZS1791	0.68314
…	…
LH38	0.61001

STEP 3: EXAMPLE OF GWAS RESULTS



Manhattan Plot Example of a Hypothetical GWAS Analysis Using Plant Height Similarity Scores

GWAS Input

Phenotypic Input

Numeric phenotype data for a trait of interest for each taxa

In this study, the trait is plant height & phenotype data includes:

- Physical measurements
- Method 1: Mean Similarity Score
- Method 2: Mean Score

Genotypic Input

Genomic marker data for each taxa

In this study using SNP datasets from:

- Mazaheri, et al. 2019
- Mural, et al. 2022

Ground Truth for Results

- Known genes or loci associated with plant height trait
- Physical height measurements

SIGNIFICANCE

- We are developing methods for recording spoken descriptions of plants in the field
- We aim to use natural language processing to determine scores for traits to perform association studies for the Wisconsin Diversity panel

NEXT STEPS

- Perform Method 1 and Method 2 for text-to-similarity score determination (STEP 2: TEXT-TO-SIMILARITY SCORE described above)
- Perform GWAS (STEP 3: EXAMPLE OF GWAS RESULTS described above) and compare to previous association studies using the Wisconsin Diversity panel

REFERENCES

Amazon Transcribe Developer Guide. (n.d.). https://docs.aws.amazon.com/pdfs/transcribe/latest/dg/transcribe-dg.pdf#what-is

Braun, I. R., Yanarella, C. F., & Lawrence-Dill, C. J. (2020). Computing on Phenotypic Descriptions for Candidate Gene Discovery and Crop Improvement. *Plant Phenomics, 2020*, 1–4. https://doi.org/10.34133/2020/1963251

Mazaheri, M., Heckwolf, M., Vaillancourt, B., Gage, J. L., Burdo, B., Heckwolf, S., Barry, K., Lipzen, A., Ribeiro, C. B., Kono, T. J. Y., Kaeppler, H. F., Spalding, E. P., Hirsch, C. N., Robin Buell, C., de Leon, N., & Kaeppler, S. M. (2019). Genome-wide association analysis of stalk biomass and anatomical traits in maize. *BMC Plant Biology*, *19*(1). https://doi.org/10.1186/s12870-019-1653-x

Mazaheri, M., Heckwolf, M., Vaillancourt, B., Gage, J. L., Burdo, B., Heckwolf, S., Barry, K., Lipzen, A., Ribeiro, C. B., Kono, T. J. Y, Kaeppler, H. F., Spalding, E. P., Hirsch, C. N., Buell, C. R., de Leon, N., & Kaeppler, S. M. (2019). *Data from: Genome-wide association analysis of stalk biomass and anatomical traits in maize*. https://doi.org/10.5061/dryad.n0m260p

Montani, I., Honnibal, M., Honnibal, M., Sofie Van Landeghem, Boyd, A., Peters, H., Paul O'Leary McCann, geovedi, jim, O'Regan, J., Maxim Samsonov, Orosz, G., Daniël de Kok, Blättermann, M., Duygu Altinok, Søren Lind Kristiansen, Kannan, M., Mitsch, R., Raphaël Bournhonesque, Edward, & Miranda, L. (2023). *explosion/spaCy: v3.5.1: spancat for multi-class labeling, fixes for textcat+transformers and more*. https://doi.org/10.5281/zenodo.7715077

Mural, R. V., Sun, G., Grzybowski, M., Tross, M. C., Jin, H., Smith, C., Newton, L., Andorf, C. M., Woodhouse, M. R., Thompson, A. M., Sigmon, B., & Schnable, J. C. (2022). Association mapping across a multitude of traits collected in diverse environments in maize. *GigaScience*, *11*. https://doi.org/10.1093/gigascience/giac080

Mural, R., Sun, G., Grzybowski, M., Tross, M. C., Jin, H., Smith, C., Newton, L., Thompson, A. M., Sigmon, B., & Schnable, J. C. (2022). *Maize_WiDiv_SAM_1051Genotype.vcf.gz genotype file*. https://doi.org/10.6084/m9.figshare.19175888.v1

ACKNOWLEDGMENTS

2021 Student Workers: Delta, Golf, Kilo, Lima, Mike, Quebec, Victor, Yankee, & Zulu

2021 WiDiv Field: Craig Abel, Jode Edwards, John Golden, Dior Kelley, Miriam Lopez, Justin Walley, & Marna Yandeau-Nelson

Helpful Discussions: Nick Lauter, Brian Dilkes, Rajdeep Khangura, & Amanpreet Kaur

Researchers Who Worked on Generating Wisconsin Diversity Panel Genotype Data: (2019) M. Mazaheri, M. Heckwolf, B. Vaillancourt, J. L. Gage, B. Burdo, S. Heckwolf, K. Barry, A. Lipzen, C. B. Ribeiro, T. J. Y. Kono, H. F. Kaeppler, E. P. Spalding, C. N. Hirsch, C. R. Buell, N. de Leon, & S. M. Kaeppler (2022) R. Mural, G. Sun, M. Grzybowski, M. C. Tross, H. Jin, C. Smith, L. Newton, A. M. Thompson, B. Sigmon, & J. C. Schnable

Please Scan to Learn More

I'M AN AGRONOMIST

IOWA STATE UNIVERSITY
Department of Agronomy

Colleen F. Yanarella [1 2], Leila Fattel [2 3], Ásrún Ý. Kristmundsdóttir [4], & Carolyn J. Lawrence-Dill [1 2 3 4 5]

[1] Interdepartmental Bioinformatics and Computational Biology, Iowa State University, Ames, IA
[2] Department of Agronomy, Iowa State University, Ames, IA
[3] Interdepartmental Genetics and Genomics, Iowa State University, Ames, IA
[4] College of Agriculture and Life Sciences, Iowa State University, Ames, IA
[5] Department of Genetics, Development, and Cell Biology, Iowa State University, Ames, IA

Funding acknowledgment: National Science Foundation (NSF), United States Department of Agriculture (USDA), Iowa State University Plant Sciences Institute, P3 NSF Research Traineeship DGE#1545463, AIIRA NSF & USDA NIFA Award #2021-67021-35329, IOW04714 Hatch Funding to Iowa State University

NSF DESE 1545453

NSF & USDA-NIFA 2021-67021-35329

IRB Study 21-179-00 Exempt Status