BMC Research Notes

DATA NOTE Open Access

Wisconsin diversity panel phenotypes: spoken descriptions of plants and supporting data

Colleen F. Yanarella[1], Leila Fattel[1], Ásrún Ý. Kristmundsdóttir[1], Miriam D. Lopez[2], Jode W. Edwards[2], Darwin A. Campbell[1], Craig A. Abel[2] and Carolyn J. Lawrence-Dill[1*]

Abstract

Objectives Phenotyping plants in a field environment can involve a variety of methods including the use of automated instruments and labor-intensive manual measurement and scoring. Researchers also collect language-based phenotypic descriptions and use controlled vocabularies and structures such as ontologies to enable computation on descriptive phenotype data, including methods to determine phenotypic similarities. In this study, spoken descriptions of plants were collected and observers were instructed to use their own vocabulary to describe plant features that were present and visible. Further, these plants were measured and scored manually as part of a larger study to investigate whether spoken plant descriptions can be used to recover known biological phenomena.

Data description Data comprise phenotypic observations of 686 accessions of the maize Wisconsin Diversity panel, and 25 positive control accessions that carry visible, dramatic phenotypes. The data include the list of accessions planted, field layout, data collection procedures, student participants' (whose personal data are protected for ethical reasons) and volunteers' observation transcripts, volunteers' audio data files, terrestrial and aerial images of the plants, Amazon Web Services method selection experimental data, and manually collected phenotypes (e.g., plant height, ear and tassel features, etc.; measurements and scores). Data were collected during the summer of 2021 at Iowa State University's Agricultural Engineering and Agronomy Research Farms.

Keywords Phenotyping, Maize, Association studies, Audio recordings, Text transcripts, Images, Wisconsin Diversity panel

Objective

Formative research using free text descriptions of plant phenotypes along with Natural Language Processing (NLP) methods has demonstrated that computing on plant phenotypes alone can recover known genotype-phenotype associations [1, 2]. Building on these successes, continued efforts to generate plant phenotype descriptions that are both structured (e.g., ontologies) and unstructured (i.e., free text) hold great promise for enabling researchers to advance analytics for phenotypes and traits, especially when these data are made publicly accessible [3].

We developed this dataset as a foundation for analyzing large volumes of spoken phenotype descriptions in a field environment. These phenotype observations were drawn from the Wisconsin Diversity panel, which contains sufficient phenotypic diversity in a field environment for various genotype-to-phenotype analyses [4–6]. Observers generating the datasets were not confined to rigid vocabularies and were not strictly limited to a list of

*Correspondence:
Carolyn J. Lawrence-Dill
triffid@iastate.edu
[1]Iowa State University, 50011 Ames, IA, USA
[2]USDA ARS, 50011 Ames, IA, USA

traits to comment on. The protocols that we have developed, along with additional pipeline development, can initiate the use of citizen scientists and become practical for the capture low-cost large-volume free-text phenotype description of plants.

Supplemental to spoken descriptions of plant phenotypes and the text derived from these observations, measurements and scores for traits of interest were also collected as ground truth. Field layout and weather data are reported, along with images of the rows in the field and aerial images from a drone. Consequently, this dataset may be useful to investigators interested in data collected from diversity panels and to those interested in processing natural language and its use in describing scientific phenomena.

The use of this dataset for investigating biological relevance and utility, including developed tools to assist in the use of spoken descriptions for field-based plant phenotype analytics is available [7].

Data description

This dataset [8] was collected and derived from observations of an experimental field at Iowa State University's Agricultural Engineering and Agronomy Research Farms in Boone, Iowa. The Wisconsin Diversity panel (686 accessions), an environmental control line (B73, the maize reference line used for genetics and genomics), and 25 positive control accessions were planted in two replicates, and observations were generated over the summer of 2021. This dataset includes the following elements (Table 1).

- Each student participant and volunteer is identified in datasets only by their code names (from the NATO phonetic alphabet).
- Audio text processing data contains the spoken data collected by the volunteers (WAV files) and descriptions of the recordings generated by student participants using Sony ICD-UX570 recorders. Additionally included are metadata (summary statistics) derived from the recordings and code to generate these statistics. Further, all intermediate files (JSON, TXT, and EXCEL files) and code to generate the final cleaned transcripts for all student participant recordings and a subset of the volunteer's recordings are included. These files provide a resource to investigators to utilize field-collected spoken natural language descriptions of maize plants.
- Methods selection data includes data and code for generating transcriptions using various Amazon Web Services (AWS) Transcribe methods. These methods include using an individualized custom vocabulary for each student participant and an example using data collected by volunteer "Whiskey" as well as a generalized custom vocabulary for each student participant and volunteer Whiskey's data, and no custom vocabulary. A subset of data was selected to process and compare to a gold standard transcription manually generated to calculate a similarity score to determine the method for transcribing all spoken descriptions collected during the summer of 2021.
- Images of each row in the field were captured with a Canon EOS Rebel T7 camera and Canon EF-S 18-55 mm Image Stabilizer Macro 0.25 m/0.8ft set to 18 mm with AF Stabilizer ON.
- Aerial still images and footage of the experimental field were captured with a Mavic 2 Pro drone by DJI.
- The field data information layout demonstrates the randomizations of the accessions planted and positive controls used. Additionally, seeds planted per row and the Iowa Phytosanitary Corn Field Inspection report conducted on the experimental field are included.
- Field prompting data includes the cards provided to student participants to prompt their behavior while collecting spoken observations in the field, information about the assigned card for each day, and logs for worker data collection. Volunteer field guide cards and logs for data collection are present. Each student participant was instructed to make three complete passes of the field, and volunteers were instructed to make one complete pass of the field. Volunteer "India" completed observations for replicate one only.
- Measurement and scoring data were collected through manual measuring and scoring by trained student participants and volunteers. Plastic measuring sticks with hash marks every 10 cm were used to measure plant height.
- Weather information was collected and reported by The Iowa Environmental Mesonet through Iowa State University [9]. Data for the weather stations nearest the Agricultural Engineering and Agronomy Research Farms for March 2021 to September 2021 and September 2020 to September 2021 are provided.

Limitations

Some audio observations were incomplete due to technical difficulties, including microphone disengagement from the recording devices or observers recording observations for the incorrect row. Also, speech-to-text pipelines and post-process cleaning steps are fallible, leading to transcription inaccuracies. These data were taken over approximately seven weeks, and there were apparent growth and developmental changes throughout the

Table 1 Overview of data files/data sets

Label	Name of data file/data set	File types (file extension)	Data repository and identifier (DOI or accession number)
2021 Wisconsin Diversity Panel Dataset	Carolyn_Lawrence_Dill_Maize_WiDiv_Summer_2021_Dataset_June_2023	Directory	CyVerse [8] (https://doi.org/10.25739/pvx4-5j31)
	/README.txt	A file with file type:.txt Provides details regarding the subdirectories.	CyVerse [8] (https://doi.org/10.25739/pvx4-5j31)
	/audio_text_processing_data	A subdirectory containing file types:.csv,.json,.py,.tar.gz,.txt,.xlsx, and.yaml Demonstrates processing of spoken observations to text files.	CyVerse [8] (https://doi.org/10.25739/pvx4-5j31)
	/aws_method_selection	A subdirectory containing file types:.csv,.json,.out,.py,.R,.txt,.wav,.xlsx, and.yaml Demonstrates methods for and accuracy of for transcribing speech-to-text using AWS.	CyVerse [8] (https://doi.org/10.25739/pvx4-5j31)
	/camera_images	A subdirectory containing file types:.pptx,.tar.gz,.txt, and.xlsx Contains images of the experimental field rows.	CyVerse [8] (https://doi.org/10.25739/pvx4-5j31)
	/drone_images	A subdirectory containing file types:.jpg,.mp4, and.txt Consists of drone-captured still images and videos of the experimental field.	CyVerse [8] (https://doi.org/10.25739/pvx4-5j31)
	/field_layout_and_seed_information	A subdirectory containing file types:.pdf,.txt, and.xlsx Includes information about taxa planted and location of taxa in the experimental field.	CyVerse [8] (https://doi.org/10.25739/pvx4-5j31)
	/field_prompting_cards_information	A subdirectory containing file types:.png,.psd,.txt, and.xlsx Demonstrates field prompting cards and directions provided to observers.	CyVerse [8] (https://doi.org/10.25739/pvx4-5j31)

Table 1 (continued)

Label	Name of data file/data set	File types (file extension)	Data repository and identifier (DOI or accession number)
	/measurement_and_scoring_data	A subdirectory containing file types:. txt and.xlsx Records of measurement and scoring data of the experimental field.	CyVerse [8] (https://doi.org/10.25739/pvx4-5j31)
	/weather_data	A subdirectory containing file types:. txt and.xlsx Reports weather data collected from stations in close proximity to the experimental field.	CyVerse [8] (https://doi.org/10.25739/pvx4-5j31)

duration of the study. Additionally, the observations within this dataset are for two replicates in the same environment, and additional years, plots, and environments could supplement these available speech data for a more robust dataset.

Abbreviations

AWS	Amazon Web Services
NLP	Natural Language Processing

Acknowledgements
We acknowledge our summer 2021 student participants, known throughout this project by their NATO code names for generating descriptive recordings and collecting scoring and measurement data. Delta, Golf, Kilo, Lima, Mike, Quebec, Victor, Yankee, and Zulu, thank you for lending your voices to this dataset and research. We thank the research groups of Nick Lauter, Marna Yandeau-Nelson, Dior Kelley, and Justin Walley for bulking the seed, making the seed available for this research, and planting. We thank Nick Lauter for his guidance on germplasm selection and his passion for working with the Wisconsin Diversity panel. We thank John Golden for assisting with seed treatment and planting. We thank Marty Sachs and the Maize Genetics Cooperation Stock Center for discussing germplasm for positive controls and making stock available for this research. We acknowledge the Predictive Plant Phenomics NSF NRT for a Program Trainee Research Materials and Supplies Grant to CFY for recording supplies, camera supplies, and lending us a drone. We thank Ashlyn Rairdin for instructing us on the use of drones for aerial images. We appreciate the Agricultural Engineering Agronomy and Central Iowa Research Farms field crew for preparing and managing the field. We appreciate the ISU Office of Research Ethics for guidance on the safety of human research subjects and for implementing their COVID-19 mitigation plan.

Author contributions
CFY and CJLD conceived this work and wrote the manuscript. All authors read, offered suggestions, and approved the final version of the manuscript. CFY, LF, ÁÝK, and CJLD are volunteers who recorded spoken observations of plants for dissemination. CJLD, DAC, and CFY managed IRB compliance, obtained IRB exemption, and coordinated student participants and individuals in the behavioral research component of this work. CFY, MDL, and JWE performed randomizations and approved the field layout design. CJLD and CFY selected the maize panel germplasm. CAA, MDL, and JWE planted the seed and approved field management practices. CFY performed image acquisition, data processing, and data organization. CFY and LF managed DOI acquisition.

Funding
This work was supported by an Iowa State University Plant Sciences Institute Faculty Scholars Award (CJLD) and the Iowa State Predictive Plant Phenomics NSF Research Traineeship (DGE-1545453; CJLD is a co-principal investigator, and CFY is a trainee). The Predictive Plant Phenomics Program provided the Program Trainee Research Materials and Supplies Grant and the Program Small Research Grant (CFY). This article was also supported by the NSF and USDA-NIFA AI Institute for Resilient Agriculture (#2021-67021-35329) to CJLD and supporting CFY and LF. This article is also a product of the Iowa Agriculture and Home Economics Experiment Station, Ames, Iowa, Project No. IOW04714 (CJLD) which is supported by USDA/NIFA and State of Iowa funds. This research was supported in part by the US. Department of Agriculture, Agricultural Research Service (USDA ARS) Project No. 5030-21000-066-000D and 5030-21000-019-000D. Mention of trade names or commercial products in this publication is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the U.S. Department of Agriculture. USDA is an equal opportunity provider and employer.

Data availability
The data described in this Data Note can be freely and openly accessed on CyVerse under Digital Object Identifiers (DOI) https://doi.org/10.25739/pvx4-5j31. Please see Table 1 and the references list for details and links to the data.

Declarations

Ethics approval and consent to participate
The ethics approval for this study was waived by Iowa State University (ISU), Ames, Iowa Internal Review Board (IRB; Study Number 21-179-00). Informed consent was obtained from all participants involved in the behavioral component of this study. All participants' data remains de-identified, and their audio recordings cannot be released as ISU IRB requires for Study 21-179-00. All volunteer observers willingly consented to participate in spoken data collection. All experiments were performed in accordance with relevant guidelines and regulations (such as the Declaration of Helsinki).

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 3 August 2023 / Accepted: 11 January 2024
Published online: 23 January 2024

References

1. Oellrich A, Walls RL, Cannon EK, Cannon SB, Cooper L, Gardiner J, et al. An ontology approach to comparative phenomics in plants. Plant Methods. 2015;11(1). https://doi.org/10.1186/s13007-015-0053-y.
2. Braun IR, Lawrence-Dill CJ. Automated methods enable direct computation on phenotypic descriptions for novel candidate Gene Prediction. Front Plant Sci. 2020;10. https://doi.org/10.3389/fpls.2019.01629.
3. Braun IR, Bassham DC, Lawrence-Dill CJ. The case for retaining natural language descriptions of phenotypes in plant databases and a web application as proof of concept. bioRxiv. 2021. https://doi.org/10.1101/2021.02.04.429796.
4. Hansey CN, Johnson JM, Sekhon RS, Kaeppler SM, Leon N. Genetic diversity of a maize association population with restricted phenology. Crop Sci. 2011;51(2):704–15. https://doi.org/10.2135/cropsci2010.03.0178.
5. Hirsch CN, Foerster JM, Johnson JM, Sekhon RS, Muttoni G, Vaillancourt B, et al. Insights into the maize pan-genome and pan-transcriptome. Plant Cell. 2014;26(1):121–35. https://doi.org/10.1105/tpc.113.119982.
6. Mazaheri M, Heckwolf M, Vaillancourt B, Gage JL, Burdo B, Heckwolf S, et al. Genome-wide association analysis of stalk biomass and anatomical traits in maize. BMC Plant Biol. 2019;19(1). https://doi.org/10.1186/s12870-019-1653-x.
7. Yanarella CF, Fattel L, Lawrence-Dill CJ, GWAS from spoken phenotypic descriptions: a proof of concept from maize field studies. bioRxiv. 2023. https://doi.org/10.1101/2023.12.11.570820
8. Lawrence-Dill CJ, Carolyn_Lawrence_Dill_Maize_WiDiv_Summer_2021_Dataset_June_2023. CyVerse Data Commons. 2023. https://doi.org/10.25739/pvx4-5j31.
9. Herzmann D. Iowa Environmental Mesonet. Available from: https://mesonet.agron.iastate.edu/.

Publisher's Note