00960222: Language, Computation and Cognition Project Guidelines

5 May 2025

Students enrolled in 00960222 are required to complete a class project which constitutes 65% of the course grade. The project is due **July 10** at 23:55. You will work on the project in pairs. In exceptional cases, upon TA approval you may work on the project individually. You are expected to work on the project continuously throughout the semester. As a rule of thumb, a good project requires at least 50 hours of work from each project member, so you should plan on dedicating at least 6 hours per week for the project during the last two months of the semester. The project deliverables are (1) a project report and (2) a link to a Github repository with your code.

Structured Projects (1 and 2) By default, you are expected to carry out either project 1 or project 2 proposed below. These projects are based on assignments from homeworks 2 and 3, respectively. Both projects consist of two parts - structured tasks and an open-ended task. The open ended task is the primary component of the project.

Open Projects (for PhD students) PhD students whose research is in NLP, psycholinguistics or related areas can (but are not obligated to) do an open research project on a topic of their choosing. The project should be *clearly* related to the topics covered in the course, and will need to be approved by the course staff. If you would like to do such a project, you are asked to submit a 1-page project proposal by May 12 and meet with the course staff.

Project Reports All projects should include a written report, which should follow from the content of the project. The report length should be 6-9 pages, roughly comparable to a 6 pages proceedings paper for Cognitive Science – the Cognitive Science Society conference or 8 pages ACL – Association for Computational Linguistics proceedings paper. Please use the Cognitive Science[1] or ACL[2] conference templates for your writeup.

The report is an integral component of the project, and *part of the grade will be given based on the quality and the clarity of the report.* You should think of the report as a scientific paper that you are submitting to a scientific conference like ACL or CogSci, to be evaluated via peer review. You can expect the reviewer to be generally literate in computational psycholinguistics, but you cannot rely on them having any specific knowledge of your work.

[1] https://cognitivesciencesociety.org/submissions/

[2] https://github.com/acl-org/acl-style-files?tab=readme-ov-file

Report requirements and recommendations:

- Report sections:
  - Abstract: a short, one paragraph summary of the main scientific and/or engineering questions that the work addresses and the main conclusions. At the end of the abstract you should have a *footnote with a link to a GitHub repository with your code*.
  - Introduction: A more elaborate exposition of the scientific background, key relevant prior work with citations, your research question(s) and why it is novel and/or interesting and your main findings.
  - Data: A short description of the dataset(s) you worked with, with key statistics such as the number of words in a textual corpus, number of participants in a human subjects experiment, etc.
  - Experiments and Results: This is the heart of the report which should describe in detail (i) *what you did*, (ii) *how you did it*, and (iii) *the results of your experiments / analyses and what we learned from them*. This section may be divided into multiple parts as you see fit, and should constitute most of the writeup. If you are using an approach that relies on previously published papers or are using code from published papers, be sure to cite them. Figures and tables are generally helpful to include in this part of the report. Figure axes should be clearly labeled and a legend should be included when relevant. Each figure and each table should include an informative caption that explains it and briefly states the conclusion(s) that can be drawn from it. Information that appears in captions can be reiterated in the text.
  - In the structured projects (1 and 2), Experiments and Results should have two subsections. (1) Structured Tasks (2) Open Ended Task. Most of the content should be dedicated to the open ended task.
  - Discussion and Conclusions: This section should summarize the key results and discuss their theoretical and/or practical implications, as well as the limitations of the work. Directions for future work may also be included in this section.
  - Bibliography
- Failure to include the sections above and their respective content will lead to deduction of points.
- It is important that your report is well structured and *clearly written*. While we will not deduct points for English grammar mistakes, if the exposition of the work is sloppy, unclear or partial, this will lead to deduction of points.
- You may include an appendix with additional figures and tables if needed. Appendix tables and figures should have captions and should be mentioned and referenced in

the main text. The appendix can only contain supplementary materials which are not strictly required for understanding your work. Any results that are key to understanding your work and its contributions should appear in the main text.

- While we ask that you link to the project's GitHub repository, you cannot expect the reader to consult your raw code in understanding your writeup.
- Before writing the report, we recommend going through a few papers from the proceedings of ACL and CogSci in order to see the exposition style and content type expected in your report.
- We expect that you *proofread* the report before submission.
- We highly recommend starting to work on the report early, at least a month before the submission deadline.

Grading The expectations from all the projects are *high*, with the following rule-of-thumb division to grade categories:

96-100	Outstanding project with in-depth work that can form the basis for an international conference paper (e.g. ACL / SCiL / CogSci).
90-95	Excellent project with solid work that can form the basis of an international workshop paper (e.g. HSP / CMCL).
80-89	A very good project with non-trivial analyses and adequate scope.
70-79	A good project with notable quality issues and/or limited scope.
60-69	A project with major quality issues and/or very limited scope.
55-59	A project that meets the minimum requirements.

Note that the project type (Structured/Open) has *no impact on the grade*. Further note that negative experimental results are just as good as positive results; grading will be based only on the quality of the ideas, the level of execution in the analyses, the scope of the project and the quality of the report.

Timeline

- **by May 12 (23:55)** team up with a partner. Register your team on Moodle and indicate the project number you are planning to do.
  - Structured Projects (1 or 2): No further action is required.
  - Open Projects: note that this option is open only to PhD students. Send us (Yevgeni, Omer and Keren) an email attaching a 1-page project proposal. In your email, include your availability for a 20 minutes meeting during this week (May 5th till May 9th) or on May 14 to discuss project details.
- **by June 11 (23:55)** all projects: submit on Moodle a 1 page progress report. For structured projects, it should specify your progress on the structured tasks, and a concrete plan for the open-ended task. Structured projects in good standing should have the structured tasks completed, and ideally have preliminary results for the open-ended task. Additionally, you will be asked to specify your availability for a 15 minutes progress update meeting on June 16th.
- **by July 10 (23:55)** Submit the final report on Moodle.

Late Submissions Policy

Teams that will not submit the June 11 progress report will be deducted 4 points from the project final grade. Late submissions will be deducted 1 project final grade point per day (up to 4 days). Final project reports submitted after the deadline will be deducted 3 points per day, up to a maximum of 5 days. Projects submitted over five days after the deadline will not be graded.

Project 1: Surprisal and RTs

In Homework Assignment 2 you examined the main result from Smith and Levy (2013) regarding the relation between reading times and surprisal using n-gram surprisals that we provided. In this project you will extend this work in a number of ways.

Structured Tasks (40 points)

Task 1: Comparison of n-gram and neural language models Download the OneStop Eye Movements *Ordinary Reading* dataset (Berzak et al., 2025), available at: https://osf.io/zn9sq, with documentation at https://github.com/lacclab/OneStop-Eye-Movements. Use the `ia_Paragraph.csv` file, which contains word-level reading time measures over paragraph.

Compute surprisal values for the words from two models: A smoothed trigram model and a neural transformer model. For the trigram model, you can use the `KenLM` package. For

the neural model, use the Pythia 70M language model (Biderman et al., 2023). You may use either the `minicons` or `text-metrics` package to extract surprisals from Pythia 70M.

Next, let's zoom in on the difference between the surprisals derived from the two models. In the following analyses it is sufficient to use a univariate model without control predictors.

1. Examine the relation between surprisal and reading times for both models. Use Total Fixation Duration (coded as IA_DWELL_TIME) as the target reading measure. Which of the two models has surprisal estimates that correlate better with human reading times?

2. Plot the relationship between the n-gram model's surprisal estimate for a word and the neural model's estimate. Each point in the graph should correspond to a single word. Describe what you see in this graph. Are the models generally well matched? On what parts of the surprisal spectrum do they disagree?

3. Pick specific interesting points from this graph (for example, points where the two models have very different surprisal estimates) and report the sentences containing the corresponding tokens. Why do you think the models should disagree?

4. Examine **spillover** in both models: look at the relationship between word probability and the *next* word's reading time (a "spillover" effect). Is the effect similar as on the current word? Bigger? Smaller? Different shape? Are the spillover effects different across the two models?

Task 2: GAM model Use the Pythia-70M estimates to fit and plot the RT-surprisal curve using a General Additive Model (GAM). Differently from the analyses above, the model should include control variables for log-frequency and word length. Examine both current word and spillover effects. Perform the GAM RT-surprisal analysis with 2–3 other reading time measures (e.g. Gaze Duration; see guide on Moodle), and compare the results across the measures. Comment on the similarities and differences that you find between the different measures.

Open-Ended Task (60 points)

Devise and carry out an additional substantial analysis (or analyses) not listed above. Your ideas may address any theoretical question or modeling challenge concerned with reading times and surprisal, or related topics. You are required to use one of the resources listed below, which provide opportunities for exploring a variety of directions, including questions related to different types of reading interactions, reader populations (e.g. L2 readers), languages, methodologies, etc.

Resources

Tools

- `KenLM`: Toolkit for training n-gram models. See usage guide here.
- `Hugging Face`: Repository of pre-trained language models.

Datasets

- **Eyetracking**
  - OneStop: Documentation. Note that OneStop includes in addition to ordinary reading, also information-seeking reading and repeated reading sub-corpora.
  - GECO (Cop et al., 2017), MECO L1 (Siegelman et al., 2022), MECO L2 (Kuperman et al., 2023), CELER (Download) (Berzak et al., 2022), PROVO (Eyetracking + Cloze) (Luke & Christianson, 2018).
  - Additional eyetracking datasets, including in languages other than English can be found here.
- **Self Paced Reading** Natural Stories SPR (Futrell et al., 2021)
- **Maze** Natural Stories Maze (Boyce & Levy, 2023)
- **EEG** UCL Corpus (Frank et al., 2015). Note that this corpus also has Eyetracking and SPR data (ESM1) (Frank et al., 2013).

Project 2 is on the next page.

Project 2: Word Embeddings and the Brain

In Homework Assignment 3 you replicated Analysis 1 from Pereira et al., 2018 on decoding words from fMRI data. In this project you will extend this work as follows.

Structured Tasks

Task 1: Sentence decoding

- Perform the analysis of Homework Assignment 3 question 3 using another type of static word embeddings (e.g. Word2vec) and compare the results to those you obtained with GloVe.
- Read Pereira et al., 2018 and describe the similarities and differences between analyzes 1, 2, and 3 in that paper.
- Use the GloVe based decoder model you trained in Homework Assignment 3 question 3 and test it on the datasets from analyses 2 & 3. Each dataset contains sentence representations (i.e. a vector representation averaged over all the words in the sentence) and the corresponding neural data from an individual subject (384 sentences from analysis 2 and 243 from analysis 3; The datasets are available in a Google Drive Folder[3]). For each dataset, use the learned decoder model to decode sentence representations and evaluate the performance via the rank accuracy method (as you did in HW3).
- Each sentence, in both datasets, is related to a specific passage (a single passage contains 3 or 4 sentences), and every passage is related to a specific broad topic (e.g., musical instrument, animals, etc. The labels for the sentences/passages are available in the Google Drive folder as well). You will need to analyze the accuracy scores from the previous section and try to identify the topics where the decoder was more / less successful in predicting the sentences.

Task 2: Sentence representations Train a decoder model on either the dataset from analysis 2 (384 sentences) or from analysis 3 (243 sentences) using both (1) the sentence representations that were used in the paper (the same representations from the structured task) and (2) sentence representations as extracted from a contextualized word embedding model (such as BERT, GPT2, GPT3, etc.). Report and compare the results from both methods.

Task 3: Brain encoder model Instead of predicting sentence identities using neural signals (i.e., neural **decoding**), you will try to predict human neural signals from the embedding vectors representations of the sentences (neural **encoding**; you can read about neural encoder in Huth et al., 2016's paper). We ask you to fit a separate linear-regression

[3] https://drive.google.com/drive/folders/1cwciPYnnmPEReE0tpX78SQlqlwL88V8b?usp=sharing

model for each voxel in the dataset related to analysis 2 (384 sentences) or 3 (243 sentences) of Pereira et al., 2018 (180 concepts). For each voxel/model, calculate the R^2 score and examine how many voxels are *significantly* associated with the information embedded in the word vectors, and how well those voxels are predicted. This analysis should be run twice: once using the non contextualized vector representations (The original vector representations from the paper), and another time, using the contextualized representations you extracted before.

Open-Ended Task (60 Points)

Go beyond the analyses presented in Pereira et al., 2018 by carrying out an additional, substantial analysis (or analyses) of your own design. You are strongly encouraged to use the fMRI dataset from Tuckute et al., 2024[4] for this task, as it offers varied opportunities for exploring the relationships between language and brain activity, which go beyond what is possible with the Pereira et al. 2018 data. Alternative resources that can be used are listed below.

Resources

Tools

- HuggingFace: Easy access to a large variety of NLP models.

Additional fMRI Datasets

- Natural Stories (Shain et al., 2020).
- Tang et al., 2023 : Data and Code.

References

Berzak, Y., Malmaud, J., Shubi, O., Meiri, Y., Lion, E., & Levy, R. (2025). Onestop: A 360-participant english eye tracking dataset with different reading regimes. *PsyArXiv preprint*.

Berzak, Y., Nakamura, C., Smith, A., Weng, E., Katz, B., Flynn, S., & Levy, R. (2022). Celer: A 365-participant corpus of eye movements in l1 and l2 english reading. *Open Mind*, *6*, 41–50.

Biderman, S., Schoelkopf, H., Anthony, Q. G., Bradley, H., O'Brien, K., Hallahan, E., Khan, M. A., Purohit, S., Prashanth, U. S., Raff, E., et al. (2023). Pythia: A suite for analyzing large language models across training and scaling. *International Conference on Machine Learning*, 2397–2430.

[4] https://github.com/gretatuckute/drive_suppress_brains

Boyce, V., & Levy, R. (2023). A-maze of natural stories: Comprehension and surprisal in the maze task. *Glossa Psycholinguistics*, *2*(1).

Cop, U., Dirix, N., Drieghe, D., & Duyck, W. (2017). Presenting geco: An eyetracking corpus of monolingual and bilingual sentence reading. *Behavior research methods*, *49*, 602–615.

Frank, S. L., Fernandez Monsalve, I., Thompson, R. L., & Vigliocco, G. (2013). Reading time data for evaluating broad-coverage models of english sentence processing. *Behavior research methods*, *45*, 1182–1190.

Frank, S. L., Otten, L. J., Galli, G., & Vigliocco, G. (2015). The erp response to the amount of information conveyed by words in sentences. *Brain and language*, *140*, 1–11.

Futrell, R., Gibson, E., Tily, H. J., Blank, I., Vishnevetsky, A., Piantadosi, S. T., & Fedorenko, E. (2021). The natural stories corpus: A reading-time corpus of english texts containing rare syntactic constructions. *Language Resources and Evaluation*, *55*, 63–77.

Huth, A. G., De Heer, W. A., Griffiths, T. L., Theunissen, F. E., & Gallant, J. L. (2016). Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, *532*(7600), 453–458.

Kuperman, V., Siegelman, N., Schroeder, S., Acartürk, C., Alexeeva, S., Amenta, S., Bertram, R., Bonandrini, R., Brysbaert, M., Chernova, D., et al. (2023). Text reading in english as a second language: Evidence from the multilingual eye-movements corpus. *Studies in Second Language Acquisition*, *45*(1), 3–37.

Luke, S. G., & Christianson, K. (2018). The provo corpus: A large eye-tracking corpus with predictability norms. *Behavior research methods*, *50*, 826–833.

Pereira, F., Lou, B., Pritchett, B., Ritter, S., Gershman, S. J., Kanwisher, N., Botvinick, M., & Fedorenko, E. (2018). Toward a universal decoder of linguistic meaning from brain activation. *Nature communications*, *9*(1), 1–13.

Shain, C., Blank, I. A., van Schijndel, M., Schuler, W., & Fedorenko, E. (2020). Fmri reveals language-specific predictive coding during naturalistic sentence comprehension. *Neuropsychologia*, *138*, 107307.

Siegelman, N., Schroeder, S., Acartürk, C., Ahn, H.-D., Alexeeva, S., Amenta, S., Bertram, R., Bonandrini, R., Brysbaert, M., Chernova, D., et al. (2022). Expanding horizons of cross-linguistic research on reading: The multilingual eye-movement corpus (meco). *Behavior research methods*, 1–21.

Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, *128*(3), 302–319.

Tang, J., LeBel, A., Jain, S., & Huth, A. G. (2023). Semantic reconstruction of continuous language from non-invasive brain recordings. *Nature Neuroscience*, 2022–09.

Tuckute, G., Sathe, A., Srikant, S., Taliaferro, M., Wang, M., Schrimpf, M., Kay, K., & Fedorenko, E. (2024). Driving and suppressing the human language network using large language models. *Nature Human Behaviour*, *8*(3), 544–561.