**Project 007 Collaboration Agreement**

This agreement describes how your work on the Psychological Science Accelerator project **007 SPAM-L** will be credited.

All projects should comply with the ten central policies of the Psychological Science Accelerator (see [Moshontz et al., 2018](https://journals.sagepub.com/doi/full/10.1177/2515245918797607); see also [the Project Monitoring policy draft](https://docs.google.com/document/d/1zLNMy7bDbWNrKF-eB3boJMQ83uHyN75XSovbsRPiUVw/edit#heading=h.ts07g58je2r5)). Decisions about deviations from these policies will be made by the project lead team, in consultation with the administrative team for this project.

Please use the following table of contents to navigate this document.

[Contact Guidelines](#_heading=h.wvb07eycpcda)

[Authorship Guidelines](#_heading=h.okvqbxxsqkvw)

[Project Descriptions](#_heading=h.u0oixcsqpy0a)

[Conduct](#_heading=h.g71jy6icvj72)

[Press](#_heading=h.h8jim8mb7fv8)

[Changes to this Collaboration Agreement](#_heading=h.16a3nl36cxg5)

[Contact](#_heading=h.lb6ipf3jvt08)

# Contact Guidelines

For the SPAML and related projects, we will be using the Canvas system <https://canvas.psysciacc.org/login/canvas> to track your contributions and send you messages. You should keep your information (email) up to date within this system, so we can contact you about the project. We will use this system to collect authorship information. You can email the team at [007spaml@gmail.com](mailto:007spaml@gmail.com).

# Authorship Guidelines

**Roles and responsibilities**

Below is information about specific project roles and their respective responsibilities. Unless you have been told otherwise, you are likely receiving this document because we are asking you to join as a data collection team. Any changes to your role on the project should be discussed with the proposing authors and documented, as this has implications for authorship ordering (e.g., you sign up to be a data collection team but assist with data analysis).

Proposing Team

* Formulating the research goals and aims
* Developing the study methodology
* Developing study materials (e.g., questionnaires and surveys)
* Assisting administrative team with the coordination of material translation
* Writing the original manuscript draft
* Incorporating feedback and developing revised manuscript draft
* Analyzing data and (in collaboration with administrative team) coordinating external code review
* Coordinating manuscript submission to journals and pre-print services
* Coordinating regular meetings with administrative team

Data Collection Teams

* Documenting ethics approval for data collection submitted in Canvas
* Collecting data from at least 50 participants and/or helping with data processing (see specific project descriptions below)
* Making reasonable efforts to check and ensure data quality (e.g., by documenting issues reported by participants)
* Reporting issues to the administrative team
* Reviewing and (if desired) providing feedback on the original and revised manuscript
* [Optional] assisting with materials translations
* Respond to project-related correspondence regularly and within a reasonable amount of time (approximately 1 week)

Administrative Team

* Leading the coordination of materials translation
* Leading the organization of lab IRB ethics approval documents
* Leading the organization of collaborator information (e.g., roles, contact information, authorship information)
* Leading the organization
* Actively maintaining a project-specific email address and answering project-related questions
* Ensuring that PSA policies are followed

***How do I become an author on a paper that results from this project?***

We anticipate that this project will result in 1-3 pre-project papers and 1 main paper (however, see information below about publishing waves). The preproject papers include the necessary work to create the appropriate stimuli for the main project. The main project paper will constitute a data paper of semantic priming response latencies and related cognitive variables. We will adhere to the [CRediT taxonomy](https://www.cell.com/pb/assets/raw/shared/guidelines/CRediT-taxonomy.pdf) to determine what counts as a “contribution”. Anyone who contributes to the project via data collection, project leadership, administration, or through some other role outlined in the taxonomy will have the chance to become an author on these papers. If you are asked or would like to make a contribution not otherwise covered in this agreement please contact [007spaml@gmail.com](mailto:007spaml@gmail.com). **Specific guidelines for each paper are outlined below.**

We anticipate that other opportunities may arise within this project. We agree to contact you through the Canvas system with collaboration opportunities and update this document accordingly.

*If you are contributing to* ***data collection*** for the main paper**,** you will be sent a form to collect authorship information and invited to be an author on the first submission of the project writeup if *(and only if)* you have fully documented your IRB submission using the Canvas system.

**Publication Waves**. The goals of the main paper include collecting data across many languages. The PSA’s involvement was targeted at publishing the first dataset with five languages reaching target sample size goals (specific languages will depend on availability of labs). At this time, we will put together the paper for publication with **all contributing labs regardless of their language**. The lead author team will then continue pursuing data collection opportunities for non-finished languages through grant funding and other agreements. As each subsequent language is completed, we will include **labs who contributed to those languages or subsequent data collection** to any publication report. Note: the updating dataset doi described below will always include all authors.

***How do we handle dissenting opinions about scientific products?***

In large-scale collaborations, disagreements are inevitable. Proposing Team members are expected to carefully consider dissenting opinions from collaborators and, to whatever extent is reasonable, discuss the issue with the dissenter. Final decisions about the scientific product, nevertheless, will be left to the Proposing Team.

***How do we handle authorship on secondary papers for this project?***

The goals of the preprojects and main paper include publishing large datasets for use by other researchers. We do not expect authorship on papers making use of these datasets, however, the authors of these papers should include a citation to the dataset they used for their analyses. During data collection, we anticipate creating a system that tracks updates to the data (see Zenodo for example, which creates dois that update with releases). Therefore, it is possible that some individuals can use data for a secondary paper before the release of the full data or publication of the main data paper. These papers should cite the version of the dataset they use for analysis through the Zenodo doi, which each contributing lab will be included for authorship. The *data* may not be separately published without the lead team permission; however, we encourage use of the data to answer theoretical or exploratory questions at any time. For example, if researchers wish to analyze the amount of semantic priming in Esperanto, they may use the Esperanto dataset for their analysis but they cannot claim to have created this dataset and “publish” it in any article, even if they contributed to the data collection. In other words, the paper describing the creation and publication of the datasets is reserved for all contributors and will be led by Erin Buchanan and the SPAM-L team. If you have questions about a publication, please contact us at [007spaml@gmail.com](mailto:007spaml@gmail.com).

***What constitutes authorship? (GLOBAL DESCRIPTION)***

All authors on a particular paper must make contributions to writing - most commonly through review and editing. A “contribution” to this category can be as small as reading the manuscript and approving it. In addition, all authors must make contributions to at least one other category of the [CRediT taxonomy](https://www.cell.com/pb/assets/raw/shared/guidelines/CRediT-taxonomy.pdf). Deviations to this rule may be made at the discretion of the project leadership team. Most authors will contribute to “Investigation” (a category which includes contributions to data collection) and “Writing - Review & Editing.” If one of the papers is a Registered Report, people contributing to data collection (and thus the “Investigation” category) will receive authorship based on the assumption that they fulfill their data collection obligations.

Principal Investigators (PIs) of contributing labs are responsible for their students and staff working on the project. This includes: honestly reporting the contributions of their lab members, evaluating whether contributions merit authorship according to the above paragraph and the CRediT table, and verifying that contributions are correctly described in any formal publication. PIs are also responsible for showing this agreement to lab members who will be authors and making sure they agree to it.

Author contributions will be reported on research products using the [CRediT taxonomy.](https://www.cell.com/pb/assets/raw/shared/guidelines/CRediT-taxonomy.pdf) The [Tenzing app](https://osf.io/preprints/metaarxiv/b6ywe/) may be a helpful tool to collect this contributorship information.

***How will authorship order be determined?***

Authorship order will be organized into the following three sections, with order determined as listed below:

*Section 1* *(Proposing Team and Admin Team)*. Main proposer listed first, followed by the admin team for the project. The remaining order of proposing authors will be determined by contribution to each specific project (e.g., not all proposers are committed to preprojects, but should be noted for their conceptualization assistance). If proposers feel they each have equal contribution, their order will be randomized, and a note describing this randomization procedure will be added to the author notes.

*Special* (*Co-leads for preprojects).* As part of encouraging more collaboration, we have invited PSA members to actively contribute to some preprojects which would normally be considered as part of the proposing teams (i.e., running complex analyses, writing portions of the paper). As part of the paper generation, these individuals will be included in Section 1 with discussion of authorship order amongst the team.

*Section 2* (*Contributions beyond data collection*). We acknowledge that there are many ways to contribute to projects. This section will include members who assist with more CRediT contributions than data collection and writing.

*Section 3 (Data collection or processing team)*. Order determined at random.

*Section 3 (PSA Admin)*. The current Director of the PSA will be listed last. This project is likely to span several Directors, and each person will be given a choice if they would like to be included in this category.

***What if I have questions or concerns related to authorship?***

We encourage you to ask questions and raise concerns about authorship as early as possible. We suggest directing questions or concerns to any [Director or Associate Directors of the PSA](https://psysciacc.org/people/) if they are not resolved by contacting the leadership team at [007spaml@gmail.com](mailto:007spaml@gmail.com).

# Project Descriptions

**Power to Stimuli:** The current focus on replication and reproducibility has driven a need to adequately power studies through appropriate sample size planning. However, estimating power and sample size is usually driven by the choice of hypothesis test and research design. An incredible number of psycholinguistic database norms have been published in the last ten years, and the issues of power and sample size have largely been ignored, as these studies do not use hypothesis testing as a main objective. This project will address how to use accuracy in parameter estimation (AIPE) and qualitative coverage approaches to determine the appropriate number of participants for data collection in a study with no hypothesis test. Data from English feature production norms, the English Lexicon Project, and participant ratings (i.e., valence, concreteness, etc.) will be used to demonstrate how to estimate variable sample sizes by item for both qualitative (feature production norms) and quantitative (priming, lexical decision tasks, judgment tasks) type data.

Watch our Presentation: [Psychonomics 2020 Presentation](https://www.youtube.com/watch?v=80B_Mxk4BpQ)

**Subtitles to STRUDEL:** A critical component to research in computational linguistics lies within large corpora or datasets that are available for hypothesis testing or stimuli control for experimental research. These data can be used to capture similarity between concepts, and one way to define similarity is the overlap between corresponding features of two or more concepts. Semantic feature production norms are created by asking participants to list the necessary and defining features of a target concept (i.e., zebra-stripes, dog-barks). However, the time and effort required to create a comprehensive set of concept-feature norms from participants is often prohibitive. This project will explore the application of the computational model STRUDEL (structured dimension extraction and labeling; Baroni et al., 2010) to the Open Subtitle Corpus to create a large, cross-linguistic, set of concept-feature norms. The Open Subtitle Corpus contains fifty-one languages with at least one million tokens and concept-feature extraction via STRUDEL provides an incredible wealth of semantic data. We will demonstrate the STRUDEL model, compare the concept-feature results with human participant data, and comment on the usefulness of computational models that provide data that is difficult to collect.

Watch our Presentation: [SCIP 2020 Presentation](https://www.youtube.com/watch?v=R0_QYTPPmVM)

**Word2Vec to Many Languages:** The Open Subtitle Corpus includes movie subtitles from fifty plus languages creating a rich dataset for exploration of cross-linguistic similarity, language structure, and other computational topics. Using fastText (a variation on word2vec), van Paridon and Thompson (2020) provide word embeddings for fifty-five languages based on the Open Subtitle and Wikipedia Corpora. However, these models were created using the same dimensional structure for each language with seemingly no support for the choice of minimum word frequency, dimension size, or window size. Previous work by Mandera et al. (2016) suggests that English and Dutch were best conceptualized at different dimension and window sizes, suggesting that no one-size-fits-all approach may be appropriate. In this project, we will outline our approach to creating word embeddings using these corpora focusing on finding the appropriately tuned model for each language. We use the rich publication history of normed data, such as age of acquisition, valence, imageability, and concreteness, as a metric to select the best model for each language. We will demonstrate the results of several languages to explore the assumption of linguistic similarity across languages. Word embeddings, programming code, and other computational tools are provided as a package for researcher use and exploration.

Watch our Presentation: [SCIP 2020 Presentation](https://youtu.be/AyT1vx768Wg)

**Semantic Priming Across Many Languages:** Semantic priming has been studied for nearly fifty years across various experimental manipulations and theoretical frameworks. Critically, the understanding of semantic priming relies on reliable, well-studied stimuli with defined similarity values. In the last twenty years, the publication rates of normed stimuli databases and corpora (i.e., large bodies of text) has exponentially increased. Further, newer computational models of concept representation have been detailed using these databases. Using these newer models, we can define similarity between concepts to create reliable stimuli for study in semantic priming. In this project, we outline the need for a database of semantic priming values, particularly in non-English languages. We detail the process for creating a large database of priming values, from which new theories and hypotheses can be examined. Further, we describe the novel outputs that this proposal will support including a framework for determining sample size for studies of this nature.

Watch our presentation: [PSACON Presentation](https://psysciacc.org/psacon2020-videos/)

***What constitutes authorship? (PROJECT SPECIFIC)***

**As noted above, all authors must contribute the writing: review & editing part of the CRediT taxonomy.**

These specific guidelines may be updated as other needs arise. We will use Canvas to inform you of these opportunities and track your contributions.

**Power to Stimuli:** contribution options include analyses and writing assistance.

**Subtitles to STRUDEL:** contributions include spell checking at least 500 rows of data.

**Words to Many Languages:** contributions include formatting 5 datasets.

**SPAML:** contributions will appear in many forms (see section 2 of authorship tier), but data collection labs should use a suggested sample size of *N* = 50 participants.

**Conference presentations:** You will be included in project authorship for conference presentations using consortium style authorship (i.e., the proposing authors + leadership + the PSA as an author). If you contributed to the data or analyses used in that presentation, you will be added to the consortium list for that presentation. You must fill out the authorship information on Canvas to make sure your information is included correctly!

# Project Timeline

Please note that the proposed project timeline is subject to change.

| Summer 2021 | Preprojects, creation of stimuli, translation, project design, writing |
| --- | --- |
| Fall 2021 | Submission of registered report, if applicable  Lab recruitment for SPAML  Translation, labs work on IRBs |
| Late Fall 2021 | Launch data collection for translated languages  Look at lab recruitment to determine goals for first wave publication  Complete + submit preprojects papers |
| November 2021 | Psychonomics + SCiP presentations |
| January 2022 | NSF Proposal for funding for labs |
| Spring 2022 | Data collection |
| Summer 2022 | Examine data collection numbers to determine publication plan timing |

# Conduct

All projects should comply with the ten central policies of the Psychological Science Accelerator (see [Moshontz et al., 2018](https://journals.sagepub.com/doi/full/10.1177/2515245918797607); see also [the Project Monitoring policy draft](https://docs.google.com/document/d/1zLNMy7bDbWNrKF-eB3boJMQ83uHyN75XSovbsRPiUVw/edit#heading=h.ts07g58je2r5)).

Each contributor is personally accountable for the accuracy and integrity of their work on this project. This includes ensuring that (1) you are operating in accordance with your local research rules and regulations, (2) your specific research products are honest and accurate (e.g., data, translations, code), and (3) your description of your role on the project is accurate.

# Press

Initial press releases will be coordinated by the Proposing Team and PSA leadership. All press releases must (a) prominently highlight the role of the PSA collaborative network, and (b) link to PSA contacts. All collaborators are strongly encouraged to speak to the press, but are expected to (1) emphasize the highly-collaborative nature of the research project and (2) encourage the interviewer to reach out to various members of the team.

# Changes to this Collaboration Agreement

As the project evolves, minor changes may be made to this collaboration agreement. In those instances, all collaborators will be notified of the change and a list of recent changes will be listed at the top of the document.

# Contact

We encourage you to ask questions and raise concerns about project-related matters as soon as possible. You can direct all questions to [007spaml@gmail.com](mailto:007spaml@gmail.com). If you have an issue or concern that you would like addressed by PSA leadership, you can contact the [Director or Associate Directors of the PSA](https://psysciacc.org/people/).