Is priming consistent across languages? Preliminary findings from the SPAML: Semantic Priming Across Many Languages

Erin M. Buchanan & The Psychological Science Accelerator

The Psychological Science Accelerator

- The PSA is a CERN for psychological science

The Psychological Science Accelerator

- The PSA is a CERN for psychological science
- Globally distributed network of researchers with more than 1000 members in 82 countries

The Psychological Science Accelerator

- The PSA is a CERN for psychological science
- Globally distributed network of researchers with more than 1000 members in 82 countries
- Open science principles and practices

The Psychological Science Accelerator

- The PSA is a CERN for psychological science
- Globally distributed network of researchers with more than 1000 members in 82 countries
- Open science principles and practices
- PSA007: Semantic Priming Across Many Languages

Semantic Priming

- Semantic priming has a rich history in cognitive psychology

Semantic Priming

- Semantic priming has a rich history in cognitive psychology
- Semantic priming occurs when response latencies are facilitated (faster) for related word-pairs than unrelated word-pairs

Semantic Priming

- Semantic priming has a rich history in cognitive psychology
- Semantic priming occurs when response latencies are facilitated (faster) for related word-pairs than unrelated word-pairs
- Usually measured with the lexical decision or naming task

Semantic Priming

- Semantic priming has a rich history in cognitive psychology
- Semantic priming occurs when response latencies are facilitated (faster) for related word-pairs than unrelated word-pairs
- Usually measured with the lexical decision or naming task
- The Semantic Priming Project (Hutchison et al., 2013) provided priming values for 1661 English word-pairs

Semantic Priming

- **Semantic** priming replicates pretty well

Semantic Priming

- **Semantic** priming replicates pretty well
- WEIRD words

Semantic Priming

- **Semantic** priming replicates pretty well
- WEIRD words
- Single language focus or multilingual individuals

Semantic Priming

- **Semantic** priming replicates pretty well
- WEIRD words
- Single language focus or multilingual individuals
- A lack of data sets that are matched on language within one study

Semantic Priming

- **Semantic** priming replicates pretty well
- WEIRD words
- Single language focus or multilingual individuals
- A lack of data sets that are matched on language within one study
- How can we leverage the computational skills found in natural language processing with the open data publications to improve this research?

Semantic Priming

- **Semantic** priming replicates pretty well
- WEIRD words
- Single language focus or multilingual individuals
- A lack of data sets that are matched on language within one study
- How can we leverage the computational skills found in natural language processing with the open data publications to improve this research?
- Goals of of the SPAML:

Semantic Priming

- **Semantic** priming replicates pretty well
- WEIRD words
- Single language focus or multilingual individuals
- A lack of data sets that are matched on language within one study
- How can we leverage the computational skills found in natural language processing with the open data publications to improve this research?
- Goals of of the SPAML:
  - Assess semantic priming across (at least) 10 languages using matched stimuli

Semantic Priming

- **Semantic** priming replicates pretty well
- WEIRD words
- Single language focus or multilingual individuals
- A lack of data sets that are matched on language within one study
- How can we leverage the computational skills found in natural language processing with the open data publications to improve this research?
- Goals of of the SPAML:
    - Assess semantic priming across (at least) 10 languages using matched stimuli
    - Provide a large-scale data set for reuse in linguistics

Semantic Priming

- **Semantic** priming replicates pretty well
- WEIRD words
- Single language focus or multilingual individuals
- A lack of data sets that are matched on language within one study
- How can we leverage the computational skills found in natural language processing with the open data publications to improve this research?
- Goals of of the SPAML:
    - Assess semantic priming across (at least) 10 languages using matched stimuli
    - Provide a large-scale data set for reuse in linguistics
- Registered Report at *Nature Human Behaviour*

The Stimuli

- Corpus Text Data: Open Subtitles Project

The Stimuli

- Corpus Text Data: Open Subtitles Project
- Subtitles have shown to be critically useful data sets for word frequency calculation (New et al., 2007; Brysbaert & New, 2009; Keuleers et al., 2010; Cuetos et al., 2012; Van Heuven et al., 2014; Mandera et al., 2015; and more)

The Stimuli

- Corpus Text Data: Open Subtitles Project
- Subtitles have shown to be critically useful data sets for word frequency calculation (New et al., 2007; Brysbaert & New, 2009; Keuleers et al., 2010; Cuetos et al., 2012; Van Heuven et al., 2014; Mandera et al., 2015; and more)
- Freely available subtitles in 63 languages for computational analysis

The Stimuli

- Corpus Text Data: Open Subtitles Project
- Subtitles have shown to be critically useful data sets for word frequency calculation (New et al., 2007; Brysbaert & New, 2009; Keuleers et al., 2010; Cuetos et al., 2012; Van Heuven et al., 2014; Mandera et al., 2015; and more)
- Freely available subtitles in 63 languages for computational analysis
- Approximately 43 languages contain enough data to be usable for these projects

The Stimuli

- For each language:

The Stimuli

- For each language:
  - Collect the top 10,000 most frequent nouns, verbs, adjectives, and adverbs

The Stimuli

- For each language:
  - Collect the top 10,000 most frequent nouns, verbs, adjectives, and adverbs
  - Find the top five most similar words using cosine from subs2vec (van Paridon & Thompson, 2021)

The Stimuli

- For each language:
  - Collect the top 10,000 most frequent nouns, verbs, adjectives, and adverbs
  - Find the top five most similar words using cosine from subs2vec (van Paridon & Thompson, 2021)
  - Cross-reference this list across languages

The Stimuli

- For each language:
  - Collect the top 10,000 most frequent nouns, verbs, adjectives, and adverbs
  - Find the top five most similar words using cosine from subs2vec (van Paridon & Thompson, 2021)
  - Cross-reference this list across languages
  - Pick the most overlapping stimuli limiting repeats and proper names

The Stimuli

- For each language:
  - Collect the top 10,000 most frequent nouns, verbs, adjectives, and adverbs
  - Find the top five most similar words using cosine from subs2vec (van Paridon & Thompson, 2021)
  - Cross-reference this list across languages
  - Pick the most overlapping stimuli limiting repeats and proper names
  - 1000 final pairs

The Stimuli

- For each language:
    - Collect the top 10,000 most frequent nouns, verbs, adjectives, and adverbs
    - Find the top five most similar words using cosine from subs2vec (van Paridon & Thompson, 2021)
    - Cross-reference this list across languages
    - Pick the most overlapping stimuli limiting repeats and proper names
    - 1000 final pairs
- Important: driven by the language, not English translation

Nonwords and Translators

- Nonwords are generated with a Wuggy-like algorithm (Keuleers & Brysbaert, 2010)

Nonwords and Translators

- Nonwords are generated with a Wuggy-like algorithm (Keuleers & Brysbaert, 2010)
- Translators check all pairs for proper translation, form, and meaning

Nonwords and Translators

- Nonwords are generated with a Wuggy-like algorithm (Keuleers & Brysbaert, 2010)
- Translators check all pairs for proper translation, form, and meaning
- They suggest the appropriate words for retaining meaning between cue-target

Nonwords and Translators

- Nonwords are generated with a Wuggy-like algorithm (Keuleers & Brysbaert, 2010)
- Translators check all pairs for proper translation, form, and meaning
- They suggest the appropriate words for retaining meaning between cue-target
- They fix nonwords to ensure they are pronounceable, not too fake

Nonwords and Translators

- Nonwords are generated with a Wuggy-like algorithm (Keuleers & Brysbaert, 2010)
- Translators check all pairs for proper translation, form, and meaning
- They suggest the appropriate words for retaining meaning between cue-target
- They fix nonwords to ensure they are pronounceable, not too fake
- Dialects are considered and separated when appropriate

Procedure

- View a simple version: https://psa007.psysciacc.org/

Procedure

- View a simple version: https://psa007.psysciacc.org/
- Overall task:

Procedure

- View a simple version: https://psa007.psysciacc.org/
- Overall task:
    - A single stream lexical decision task

Procedure

- View a simple version: https://psa007.psysciacc.org/
- Overall task:
  - A single stream lexical decision task
  - All words cue-target are judged, cue-target linked by order

Procedure

- View a simple version: https://psa007.psysciacc.org/
- Overall task:
  - A single stream lexical decision task
  - All words cue-target are judged, cue-target linked by order
- Trials are formatted as:

Procedure

- View a simple version: https://psa007.psysciacc.org/
- Overall task:
  - A single stream lexical decision task
  - All words cue-target are judged, cue-target linked by order
- Trials are formatted as:
  - A fixation cross (+) for 500 ms

Procedure

- View a simple version: https://psa007.psysciacc.org/
- Overall task:
  - A single stream lexical decision task
  - All words cue-target are judged, cue-target linked by order
- Trials are formatted as:
  - A fixation cross (+) for 500 ms
  - CUE or TARGET in Serif font

Procedure

- View a simple version: https://psa007.psysciacc.org/
- Overall task:
  - A single stream lexical decision task
  - All words cue-target are judged, cue-target linked by order
- Trials are formatted as:
  - A fixation cross (+) for 500 ms
  - CUE or TARGET in Serif font
  - Lexical decision response (word, nonsense word)

Procedure

- View a simple version: https://psa007.psysciacc.org/
- Overall task:
  - A single stream lexical decision task
  - All words cue-target are judged, cue-target linked by order
- Trials are formatted as:
  - A fixation cross (+) for 500 ms
  - CUE or TARGET in Serif font
  - Lexical decision response (word, nonsense word)
  - Keyboards are WILD

Procedure

- View a simple version: https://psa007.psysciacc.org/
- Overall task:
    - A single stream lexical decision task
    - All words cue-target are judged, cue-target linked by order
- Trials are formatted as:
    - A fixation cross (+) for 500 ms
    - CUE or TARGET in Serif font
    - Lexical decision response (word, nonsense word)
    - Keyboards are WILD
    - 400 pairs = 800 trials

Power and Study Design

- Power focused on using accuracy in parameter estimation to adequately measure each individual item (see anything by Ken Kelley)

Power and Study Design

- Power focused on using accuracy in parameter estimation to adequately measure each individual item (see anything by Ken Kelley)
- We simulated using the English Lexicon Project and Semantic Priming Project

Power and Study Design

- Power focused on using accuracy in parameter estimation to adequately measure each individual item (see anything by Ken Kelley)
- We simulated using the English Lexicon Project and Semantic Priming Project
  - Minimum: $n = 50$ per target word by condition (related, unrelated)

Power and Study Design

- Power focused on using accuracy in parameter estimation to adequately measure each individual item (see anything by Ken Kelley)
- We simulated using the English Lexicon Project and Semantic Priming Project
  - Minimum: $n = 50$ per target word by condition (related, unrelated)
  - Stopping: $SE = .09$

Power and Study Design

- Power focused on using accuracy in parameter estimation to adequately measure each individual item (see anything by Ken Kelley)
- We simulated using the English Lexicon Project and Semantic Priming Project
  - Minimum: $n = 50$ per target word by condition (related, unrelated)
  - Stopping: $SE = .09$
  - Maximum $= n = 320$

Power and Study Design

- Power focused on using accuracy in parameter estimation to adequately measure each individual item (see anything by Ken Kelley)
- We simulated using the English Lexicon Project and Semantic Priming Project
  - Minimum: $n = 50$ per target word by condition (related, unrelated)
  - Stopping: $SE = .09$
  - Maximum $= n = 320$
- Adaptive sampling checks and samples pairs once an hour to randomize the study

Data Provided

- This procedure creates data at many levels

Data Provided

- This procedure creates data at many levels
    - Subject/trial level: for every participant

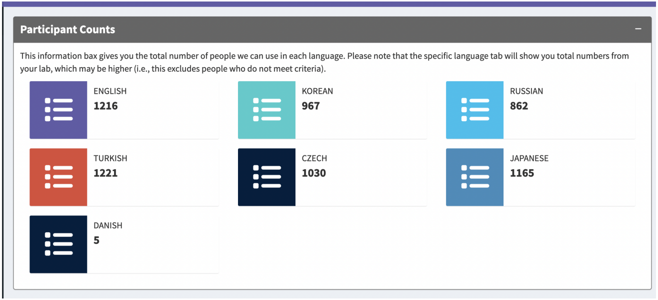
Data Provided

- This procedure creates data at many levels
  - Subject/trial level: for every participant
  - Item level: for each individual item, rather than just cue or just concept

Data Provided

- This procedure creates data at many levels
  - Subject/trial level: for every participant
  - Item level: for each individual item, rather than just cue or just concept
  - Priming level: for each related pair compared to the unrelated pair

Current Data Collection



Participant Counts

This information bax gives you the total number of people we can use in each language. Please note that the specific language tab will show you total numbers from your lab, which may be higher (i.e., this excludes people who do not meet criteria).

ENGLISH
1216

KOREAN
967

RUSSIAN
862

TURKISH
1221

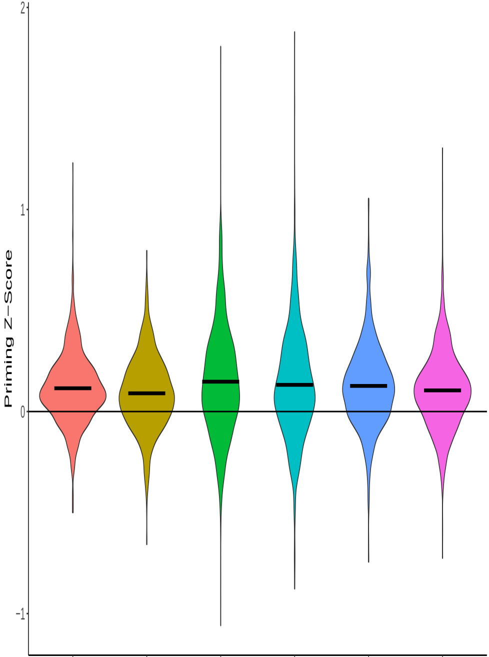
CZECH
1030

JAPANESE
1165

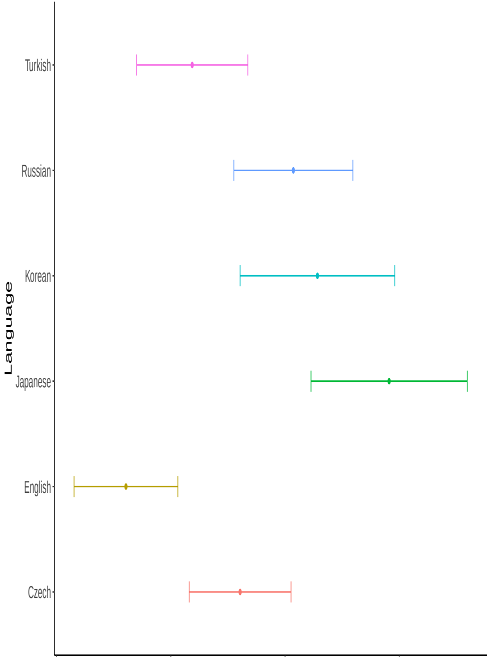
DANISH
5

*Big thanks to ZPID and Harrisburg U

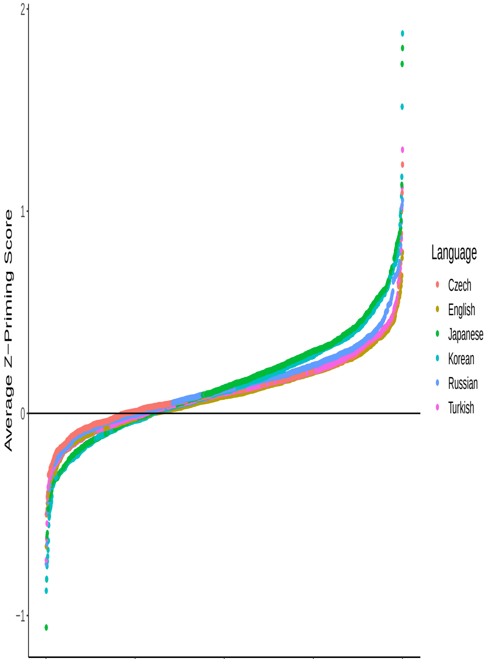
Priming Distribution Results



Priming Comparison



Item Level Priming Results



Final Thoughts

- This work to diversify participants, languages, and researchers represented is aided by big team science approaches

Final Thoughts

- This work to diversify participants, languages, and researchers represented is aided by big team science approaches
- Priming effects are found across different writing systems

Final Thoughts

- This work to diversify participants, languages, and researchers represented is aided by big team science approaches
- Priming effects are found across different writing systems
- Variability between languages appears to be approximately .02

Final Thoughts

- This work to diversify participants, languages, and researchers represented is aided by big team science approaches
- Priming effects are found across different writing systems
- Variability between languages appears to be approximately .02
- More languages currently underway

Recruitment and any Questions?

- Thank you for listening!

Recruitment and any Questions?

- Thank you for listening!
- We want you - join our team for data collection by contacting me

Recruitment and any Questions?

- Thank you for listening!
- We want you - join our team for data collection by contacting me
  - All levels of researchers welcome

Recruitment and any Questions?

- Thank you for listening!
- We want you - join our team for data collection by contacting me
  - All levels of researchers welcome
  - Authorship is provided for those who meet the collaboration agreement

Recruitment and any Questions?

- Thank you for listening!
- We want you - join our team for data collection by contacting me
  - All levels of researchers welcome
  - Authorship is provided for those who meet the collaboration agreement
- Interested in the code? Check out https://github.com/SemanticPriming/SPAML

Recruitment and any Questions?

- Thank you for listening!
- We want you - join our team for data collection by contacting me
  - All levels of researchers welcome
  - Authorship is provided for those who meet the collaboration agreement
- Interested in the code? Check out https://github.com/SemanticPriming/SPAML
- All PSA collaborators are listed with their author information online