Issues of Generalization from Unreliable or Unrepresentative Psycholinguistic Stimuli: A Case Study on Lexical Ambiguity

Jiangtian Li (jiangtian.li@utoronto.ca)
Department of Psychology, University of Toronto Scarborough, 1265 Military Trail
Toronto, Ontario, M1C 1A4, Canada

Blair C. Armstrong (blair.armstrong@utoronto.ca)
Department of Psychology, University of Toronto Scarborough, 1265 Military Trail
Toronto, Ontario, M1C 1A4, Canada

Abstract

We conducted a case study on how unreliable and/or unrepresentative stimuli in psycholinguistics research may impact the generalizability of experimental findings. Using the domain of lexical ambiguity as a foil, we analyzed 2033 unique words (6481 tokens) from 214 studies. Specifically, we examined how often studies agreed on the ambiguity types assigned to a word (i.e., homonymy, polysemy, and monosemy), and how well the words represented the populations underlying each ambiguity type. We observed far from perfect agreement in terms of how words are assigned to ambiguity types. We also observed that coverage of the populations is relatively poor and biased, leading to the use of a narrower set of words and associated properties. This raises concerns about the degree to which prior theoretical claims have strong empirical support, and offers targeted directions to improve research practices that are relevant to a broad set of domains.

Keywords: generalization crisis; sample representativeness; lexical ambiguity; semantic ambiguity; homonym; polyseme; monoseme

Introduction

Cornerstones of the scientific endeavor include developing reliable and valid procedures to draw inferences regarding hypotheses of theoretical interest. There have been extensive discussions regarding failures on this front focusing on the distinct but interrelated issues of reliability and validity. For example, the replication crisis and potential solutions thereto has garnered extensive scientific and public attention (Baker, 2016; Open Science Collaboration, 2015; Pashler & Wagenmakers, 2012), and led to a number of positive changes in research practice. Another vein of work has focused on generalization. For instance, classic work by Clark (1973) raised concerns about how to use statistics to appropriately generalize from samples of words to their underlying populations ("stimuli as fixed-effects fallacy"). This issue has recently been subsumed as part of a broader issue referred to as the "generalizability crisis" (Yarkoni, 2022). This crisis is concerned with cases wherein experimental results are replicable, but there is a mismatch between the verbal hypotheses and the statistical methods used to support an inference. Thus, "significant" effects may not generalize to the theoretical construct of interest.

A number of factors may lead to a mismatch between a verbal hypothesis and the statistical methods used to draw an inference. Our work focuses on the sample stimuli used to draw inferences, particularly in terms of whether stimuli are reliably assigned the same verbal label (i.e., terminological alignment across different studies), and whether the stimuli are representative of the populations that they supposedly represent. Concerns on these fronts may be exacerbated by the partial or complete re-use of stimulus sets across different publications (e.g., Armstrong & Plaut, 2008; Beretta et al., 2005; Rodd et al., 2002). The initial selection of experimental items, if completed manually, may also lead to the identification of biased and non-representative samples that mis-estimate population-level effects, with potentially major ramifications (Forster, 2000). Taken together, these potential issues may create an "upside down pyramid" wherein a broad theory is supported by a limited range of evidence (Frost et al., 2019).

Here, we use lexical ambiguity as a case study to probe the aforementioned issues in a broad body of research. We first identified all relevant studies (i.e., publications, which could include multiple experiments) on lexical ambiguity, including different types of ambiguity (e.g., homonym, polysemes), from a range of interdisciplinary databases and extracted their stimuli. Next, we analyzed the stimuli to answer the following research questions:

Q1: Do studies agree upon the ambiguity type that a given word represents? (terminological alignment)

Q2: Do the words represent the ambiguity types they were sampled from? (sample representativeness)

In answering these questions, we highlight problematic issues and propose targeted improvements applicable both in this specific area of research and beyond.

Methodology

Identifying Studies for Analysis

We identified studies for our analysis following the Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) guidelines (Page et al., 2021) aiming to minimize bias and oversight. Due to space constraints, we provide a brief summary of the key aspects of this procedure here. Our search covered five databases to reflect the interdisciplinary nature of lexical ambigu-

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

ity research: Scopus, APA PsychInfo, Linguistics and Language Behavior Abstracts, CogSci proceedings, and PhilPapers. We used the following keywords to search titles and abstracts for studies related to lexical ambiguity published in the cognitive sciences, broadly construed: "polysem*, homonym*, lexical ambiguit*, word sense disambig*, word sense induct*, semantic ambigu*." We also manually added a handful of studies that were not in the search but that we considered relevant. We only included studies which had been cited at least once, and thus were having some measurable impact on the literature. Collectively, this process identified 8004 studies.

The aforementioned publications were imported into the Covidence system (Veritas Health Innovation, 2023) for additional screening regarding whether the study actually related to the study of lexical ambiguity in the cognitive sciences, was a peer-reviewed article (e.g., publication of dissertations outside of journals were excluded), and for which we could access a full-text version of the article. We also used this screening to focus the scope of inclusion to lexical (semantic) ambiguity per se and excluded studies focusing on other related but distinct constructs such as homophony, puns, vagueness, and phrasal, idiomic, referential, and structural ambiguity. In total, this screening identified 1411 studies.

Of these 1411 studies, 542 were empirical, covering behavioral, neurobiological, and neuropsychological methodologies. Among them, 411 used English stimuli, with 235 providing their stimuli. To give a sense for overall publication trends, we grouped these studies into 5-year bins and plotted the number of studies in each bin in Figure 1. This plot shows (a) substantial growth in studying lexical ambiguity over the past decades, (b) that although English remains the dominant language studied, there has been a considerable increase in other languages, thus reducing Anglocentrism (Share, 2008), and (c) the proportion of studies conducted in English for which stimuli are available has increased substantially, reflective of the adoption of open science practices.

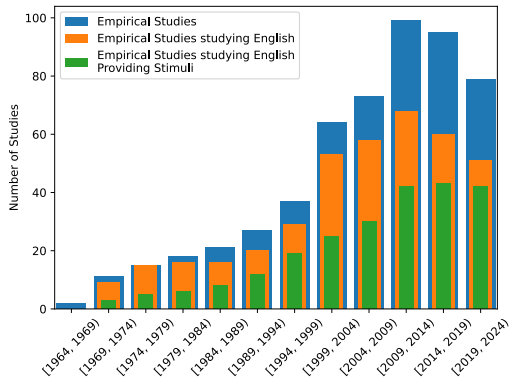

Figure 1: Number of Studies over Time

Preparing Sample Words for Analysis

Our goal was to examine the generalization of findings from homonym, polysemes, and monosemes. However, due to varying naming conventions across the diverse literature that we sampled from, we standardized the ambiguity types as labeled in studies into five standard categories. The first three category labels are the most specific and were used when studies explicitly distinguished between (related) senses and (unrelated) meanings:

(1) **polysemes** (i.e., words with related senses, but no unrelated meanings). Examples of labels: "polyseme", and "words with many senses";
(2) **homonyms** (i.e., words with unrelated meanings, but may have related senses). Examples of labels: "homonym," and "verbs with multiple meanings";
(3) **monosemes** (i.e., words with only a single sense. Examples of labels: "monoseme" and "unambiguous (neither homonym nor polyseme)."

Some (typically older) studies did not make a distinction between the aforementioned types and were labeled either as (4) **ambiguous words** (i.e., words with many meanings and/or senses) or (5) **non-homoymous words** (i.e., either monosemes or polysemes). We did not analyze these types here.

We removed other items such as nonwords, pseudowords, homophones (e.g., night/knight), filler words (i.e., words that were not analyzed in studies), and words associated with new, artificially created meanings. For each word, we also extracted the number of (unrelated) meanings (NOM) and number of (related) senses (NOS) from the Wordsmyth dictionary (Wordsmyth, 2024), which is widely used in studies of lexical ambiguity (e.g., Armstrong & Plaut, 2016; Rodd et al., 2002), and word frequency information from Brysbaert and New (2009). We removed 1912 occurrences for which either a dictionary information or frequency value was unavailable. Lastly, we only kept one unique word/ambiguity type pairing for each study, which removed 6966 duplicates (e.g., because words were re-used in multiple experiments in a study). This process resulted in 3321 unique words (types) and 12980 occurrences (tokens) across 214 studies. This dataset included 880 unique homonyms, 1368 unique polysemes, and 228 unique monosemes, for a total of 2033 words (6481 tokens) in the key types of interest. Note that some words appear in multiple ambiguity types because they were labeled differently across studies, an issue that we examine later.

Analyses

Q1: Do studies agree upon the type of ambiguity that a given word represents?

If different ambiguity types are assigned to the same word across studies, this would raise important concerns regarding whether researchers are actually mea-

suring and discussing the same construct across studies. Here, we examined how often studies agreed on an ambiguity type label for a given word. To do so, we first removed words which only appeared in a single study, leaving 1169 unique words. To gain initial insight, we then plotted a Venn diagram for the type labels to visualize their overlap (see Figure 2). This figure shows substantial amounts of disagreement in the labels used across types, with monosemes having the highest proportion of words with disagreement.

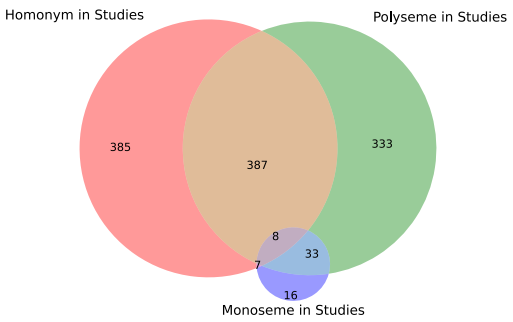

Figure 2: Venn Diagram of Word Types labeled as Homonyms, Polysemes, and Monosemes in Studies.

However, the Venn diagram only shows what labels have been given to each word type without tapping into the proportions with which each label was used for each word. next, we quantified the percentage of label agreement for each word. Specifically, we calculated its agreement percentage by dividing the number of occurrences of the most frequent label by the total number of occurrences across studies. Overall agreement was relatively high, at 88% (homonyms (H): 84%, polysemes (P): 82%, monosemes (M): 70%).[1] This recapitulates our initial insights that there is non-trivial disagreement in the labels, with particularly high amounts of disagreement for the monosemes. Given that the finer distinctions among homonyms, polysemes, and monosemes gained popularity after Rodd et al. (2002), we repeated these same analyses dividing our data into two periods: prior to 2002 and since 2002. We found that agreement since 2002 was consistently higher than prior to 2002, particularly among monosemes and polysemes (since 2002: H: 86%, P: 83%, M: 68%, total: 89%; prior to 2002: H: 84%, P: 74%, M: 54%, total: 86%). These results indicate that the agreement has improved with time but is still far from perfect. Clearly, more substantive intervention is needed to enable more consistent labeling of these stimuli and by proxy, to make consistent inferences regarding the underlying constructs. We return to this point in the discussion.

[1]Some words appear in multiple ambiguity types because they were labeled differently across studies. This leads to higher overall agreement than the average across types.

Label Agreement between Studies and a Dictionary "Gold Standard" We assessed how well the ambiguity type from the studies agreed with a "gold standard" ambiguity type label from the Wordsmyth dictionary. In the dictionary, homonyms were defined as words with more than one dictionary entry (meaning), and possibly several definitions (senses) under that entry. Polysemes were words with one entry (meaning) but more than one definition (sense). Monosemes were words with one entry and one definition. We scored every word *token* in our dataset labelled as either a homonym, polyseme, or monoseme in terms of whether it agreed (1) or disagreed (0) with the dictionary label. This yielded an overall agreement of 76%. We also calculated the percentage of correct labeling for each word *type*, which yielded a mean of 75%. The type-specific percentage was 89% for homonyms, 73% for polysemes, and 42% for monosemes. These agreement rates are similar patterns to the agreement *across* studies from the prior section.

Q2: Do the studied words represent the ambiguity types they were sampled from?

We evaluated how well words used in studies represent the population of ambiguity types in three separate subanalyses. First, we examined how much of the population of words with each ambiguity type has been sampled in studies. Here, greater coverage should increase the likelihood that findings are robust to the entire population. Next, we examined how often the same stimuli were included in multiple studies. Here, high degrees of re-use would be indicative of a bias in word sampling. We also examined whether different rates of item re-use may be explained by the frequency with which each word was used in natural language that is typically known by experimental participants (which we term "common English"). Finally, we delved into the properties underlying the ambiguity type: NOM and NOS and whether those distributions match those in common English.

What Proportion of Ambiguous Words in English are Studied? To prepare for our assessment of how well words used in studies represent the English language (defined as words that occur in the Brysbaert and New (2009) corpus and the Wordsmyth dictionary), we used a standardized ambiguity type label for each word based on the Wordsmyth dictionary, as described above. This procedure yielded 891 unique homonyms, 14647 unique polysemes, and 9777 unique monosemes. We consider this set representative of "common English" words.

We then plotted Venn diagrams to visualize how well the sampled words for each ambiguity type covered the population of words of that type in common English. Due to space constraints, we illustrate these plots with the data for homonyms used in studies and verbally summarize the observed results for the plots of the other item types. The results for homonyms used in studies are

presented in Figure 3. The left and right circles plot the standardized labels of homonyms and polysemes in Common English, respectively. The middle circle that partially overlaps with the other two denotes homonyms as labeled in studies. The studied homonyms covers more than half of the homonyms in common English (527 out of 891, 59%). However, items labeled as homonyms also include a substantial number (345) of polysemes and a few (8) monosemes in common English. Thus, current coverage of homonyms in studies is decent, but there is disagreement regarding whether a substantial minority of the words are homonyms or polysemes.

In the analogous plots for polysemes, the studied polysemes only covered a small proportion of polysemes in common English (1135 out of 13512, 8%), and included some homonyms (173) and monosemes (59). Finally, most of the items labeled as "monosemes" in studies were not in fact monosemes according to the dictionary, and they covered less than 1% of monosemes in common English (48 out of 9777). Taken together, these results indicate that except for homonyms, a relatively small proportion of common English is represented in current samples and is being used to support broad generalizations about the effects of ambiguity.

Do the sampled words represent the word frequency distributions from their underlying populations? Next, we investigated whether the samples used in studies were random representations of their underlying populations or if they were biased in some way. As a basic, intuitive check, we examined the top 10 most studied ambiguous words in each type (only the data for homonyms and polysemes are presented, see Table 1). These tables present both the frequency of occurrence of each word across studies, as well as the proportion of times that the stimuli was used across studies investigating homonyms and polysemes, respectively. These tables illustrate that there is extensive re-use of stimuli in the literature, with some homonyms in particular appearing in close to half of all studies of homonymy. The situation is less extreme for polysemes, but nevertheless many words are present in 10-20% of all studies of polysemy. Given that studies overall have only sampled 59% of homonyms and 8% of polysemes once, it is clear that these samples are biased and not representative random samples of the underlying population.

One potential explanation for the bias is that researchers are factoring word frequency (either implicitly or explicitly) into their sampling methods. For instance, more frequent words may be sampled more often, and very low frequency words may be avoided entirely out of concern that participants would not know them (Brysbaert et al., 2018). To evaluate this possibility, we first plotted the frequency of a word being studied against its (base-10, log-transformed) word frequency (hereafter lg10WF) in Figure 4. This figure hints at modest positive correlation between studied frequency and word frequency, however the relationship is clearly imperfect, with very few appearances of relatively high frequency (> 5 lg10WF) and low frequency (< 1 lg10WF) words. It also indicates that findings of monosemes are almost exclusively driven by items with frequencies < 4 lg10WF.

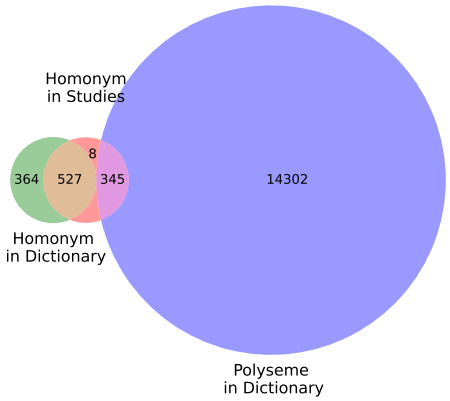

Figure 3: Venn diagram for the coverage of homonyms in studies (central circle) and in the dictionary (left circle), and polysemes in the dictionary (right circle)

Homonym	Frequency	Proportion
bat	33	.48
fan	31	.45
bank	29	.42
pen	29	.42
ring	28	.41
jam	24	.35
seal	24	.35
calf	24	.35
ball	23	.33
port	22	.32

Polyseme	Frequency	Proportion
chicken	15	.17
book	14	.16
cold	14	.16
orange	11	.13
glass	11	.13
letter	10	.12
atmosphere	10	.12
bag	9	.10
tongue	9	.10
tape	9	.10

Table 1: Top 10 Homonyms and Polysemes

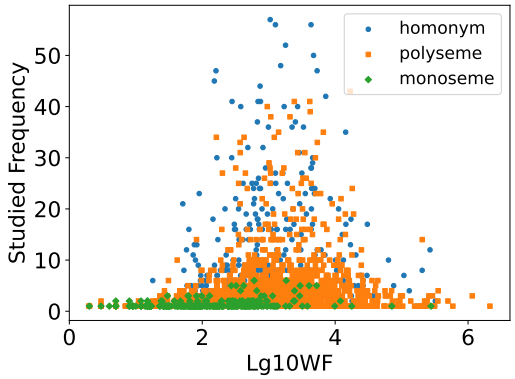

Figure 4: Scatter Plot of Studied Frequency vs. Lg10WF

To better assess the quantitative relationship between studied frequency and word frequency, we ran linear regression analyses using the log-transformed frequency

of each word to predict its studied frequency. We first ran the regression including the data from all ambiguity types and observed a very modest but significant relationship between these variables (adjusted $R^2 = .04$, $F(1, 3319) = 147.4$, $p <= .001$). We replicated this analysis separately for each ambiguity type and obtained similar results (adjusted R^2 between .05–.07). Hence, the relationship between these variables does not explain the uneven sampling rates for different words, and some other yet to be determined biases must be shaping this sampling process, which is having an as-yet unknown impact on the generalization of experimental results.

Do the sampled words represent the NOM and NOS distributions in their underlying populations? Distilling a word down to a homonym, polyseme, or monoseme is a coarse-grained distinction that may overlook the richer and more detailed underpinnings of lexical ambiguity. For instance, two polysemes could differ substantially in terms of their NOS, and these differences could potentially elicit different ambiguity effects. Thus, developing broad theories of homonymy and polysemy requires representative samples that span the range of the NOM and NOS dimensions.

To assess how well homonyms, polysemes, and monosemes are sampled from the aforementioned dimensions, we used the dictionary to derive cumulative density plots for each ambiguity type for NOM and NOS based on the words in our common English dataset. Next, we derived the analogous plots for our studied words. The results are presented in Figure 5. We corroborated our informal inferences regarding differences between the studied and common English distributions using Kolmogorov-Smirnov tests for equality across the compared distributions. Only the distribution of NOM for monosemes did not differ significantly (p=.66).

For NOM, nearly 40% of studied homonyms used in studies only have one dictionary meaning, and approximately 10% of studied polysemes have more than one meaning. Thus, if the dictionary classifications are accurate, the studied stimuli do not fully reflect the NOM distribution in natural language. Alternatively, these results could indicate that the dictionary does not accurately enumerate most word meanings. Prior normative work on homonyms, however, suggests that dictionaries such as WordSmyth tend to enumerate more meanings than are known to average language users (e.g., archaic uses of words; Armstrong, Tokowicz, & Plaut, 2012) and rarely fail to list a known meaning. This suggests that many inferences made to putative homonyms involve substantial numbers of non-homonyms.

For NOS, all three ambiguity types from studies differ from those observed in Common English. Notable differences here include that monosemes in studies tend to have several senses according to the dictionary, that both homonyms and polysemes have more senses on average

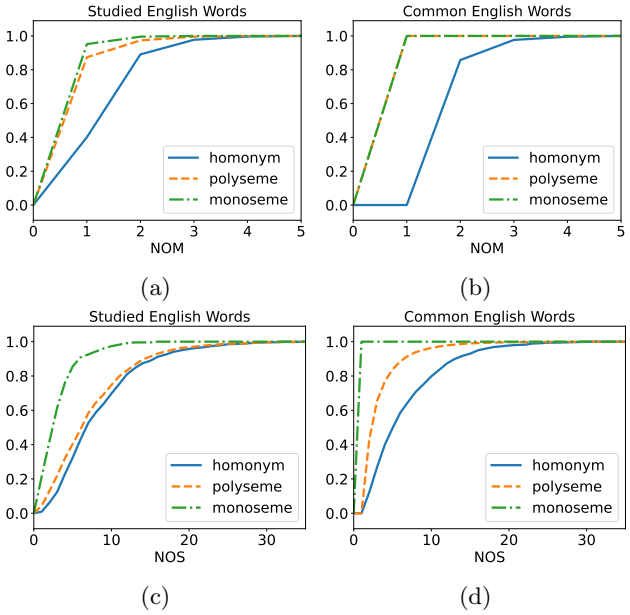

Figure 5: Cumulative distributions for NOM (top) and NOS (bottom) for studied words (left) and Comomon English words (right)

than observed in common English, and homonyms and polysemes have relatively matched NOS distributions in studies whereas in natural language homonyms tend to have more senses than polysemes. These last two differences may be the result of the experimental selection process: e.g., in classic low/high factorial designs, sampling items with higher NOS would boost the observed effects relative to a monoseme control. Nevertheless they impact how we generalize ambiguity effects to coarser verbal labels such as homonyms versus polysemes.

Discussion

We evaluated how well sample stimuli used in studies can support broad inferences regarding lexical ambiguity. Our first research question considered how consistently (reliably) researchers associate a given word with a particular ambiguity type (Q1), because issues in terminological alignment, that is, the same terminology referring to slightly different underlying constructs across studies, can lead to generalization failures. We found that there was substantial but far from perfect agreement across studies and between studies and "gold standard" labels from the dictionary. Further studies of alternative gold standards (e.g., other dictionaries) could help bolster and refine our claims, however, our preliminary work on this front suggests that these effects are not attributable to the specific dictionary we used in our work and that has been used in prior ambiguity research (e.g., Armstrong & Plaut, 2008; Rodd et al., 2002).

Our second research question considered the degree to which studied words represent the broader ambiguity

type populations and thus form a strong basis for generalization. We observed that only a small portion of polysemes and monosemes have been included in prior studies; the situation for homonyms is considerably better, but still far from ideal. This finding means that a relatively small set of items is being used to make broad generalizations about lexical ambiguity. This is an "upside down pyramid" (Frost et al., 2019) because generally a wide empirical basis should serve as the basis for specific theoretical claims. Moreover, the sampled items are clearly biased, sample NOS and NOM are not reflective of the population distributions on these dimensions, and there has been with extensive amounts of re-use of a subset of polysemes and especially of homonyms across studies. The degree of re-use is not well-explained by the word's frequency in natural language. Although there may be some pragmatic reasons to explain some of this bias (e.g., avoiding words that participants may not know; accelerating study development, replicating effects, and avoiding re-justifying stimulus selection methods during peer-review), recently reported interaction effects between frequency and certain properties of ambiguity such as NOS (Jager & Cleland, 2015) raise concerns about the impact of such bias. Researchers also almost invariably did not report the procedure used to select a sample and to define the ambiguity type population, likely relying upon manual selection of stimuli representative of a verbal hypothesis in question, an established source of biased samples (Forster, 2000).

These findings raise important concerns regarding the reliability and validity of prior theoretical inferences in the field. To be clear, our results do not indicate that prior inferences are necessarily wrong. Rather, they highlight issues with the stimuli used to support statistical inferences that reduce our confidence that prior broad verbal theoretical claims are necessarily right. Making definitive claims on this front will necessarily involve new empirical research that explicitly considers these issues.

Recommendations for future work

Our work identifies several targeted directions for improving the experimental support for theoretical inferences in the study of lexical ambiguity that are applicable to analogous issues throughout the cognitive sciences.

First, individual studies and the field more broadly must better define the population of items that a construct (e.g., homonymy) denotes to facilitate integration of insights. A better definition of the population is also essential in ensuring that samples represent this population appropriately (e.g., with respect to NOM and NOS).

Second, researchers must collectively strive to improve terminological (label) agreement. More explicit operationalization of the target construct should at least partially alleviate the sub-optimal levels of agreement. Another potential and complementary way to do so is to supplement subjective labelling procedures with a common standard (e.g., dictionary-based labels), even if that standard leaves to be desired in some respects (for discussion, see, e.g., Armstrong, Tokowicz, & Plaut, 2012). Recent computational modeling may offer alternative methods for generating labels that are also computationally explicit and less costly (in time and resources) than subjective labeling procedures, highlighting the value of interdisciplinary solutions to this issue (Li & Joanisse, 2021; Trott, 2024). When using subjective labeling procedures, it is critical that there is an explicit assessment of the reliability of this procedure, and that imperfect reliability be taken into consideration. On a related note, a potential limitation of our work is that our meta-labeling (i.e., how we distilled labels from individual studies down to the categories of homonymy, polysemy, etc.) could suffer from reliability issues. Although we consider this to be unlikely, this is fundamentally an empirical question requiring an independent re-labeling of our data. We are making our code and data available to this end.

Third, samples should be drawn from a population using an operationalized and explicitly reported procedure. Use of automated procedures (e.g., Armstrong, Watson, & Plaut, 2012) would avoid bias from manual selection (Forster, 2000), potentially speed the selection process, and facilitate the evaluation of whether a sample that satisfies the selection constraints (e.g., matching words from two ambiguity types on frequency) actually represents its population (e.g., by generating samples several times and comparing how much they overlap and cover the population). It would also facilitate using new samples versus re-using prior samples. Re-use causes broad theoretical claims to be dependent on a relatively narrow set of re-used items which, if called into question, could undermine an extensive body of research.

Fourth, larger samples of stimuli are useful to ensure that a larger portion of the population is included across studies. If data from a study are shared, this will also allow for the re-analysis of subsets of the data if issues are ever identified (e.g., inconsistent ambiguity type labeling in how to label some words), or if new potential confounds are discovered (Gernsbacher, 1984).

Finally, our work depended on the availability of the stimuli used in prior studies. Although the majority of studies provide these materials, in the age of open science, all study materials should be made available.

Conclusion

Our investigation of the word samples used to study lexical ambiguity has revealed several potential issues that impact the degree to which these stimuli form a strong basis for drawing theoretical inferences. However, consideration of these issues has provided constructive guidance for how to improve reliability, representatives, and ultimately generalizability of findings in this field that are applicable in many areas of the cognitive sciences.

Acknowledgments

This work was supported by NSERC Grant RGPIN-2017-06310 to Blair Armstrong. The authors are grateful to the many research assistants who worked in the UTSC Computation and Psycholinguistics Laboratory for their contributions.

References

Armstrong, B. C., Watson, C. E., & Plaut, D. C. (2012). Sos! an algorithm and software for the stochastic optimization of stimuli. *Behavior Research Methods*, *44*(3), 675–705. https://doi.org/10.3758/s13428-011-0182-9

Armstrong, B. C., & Plaut, D. (2008). Settling dynamics in distributed networks explain task differences in semantic ambiguity effects: Computational and behavioral evidence. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *30*.

Armstrong, B. C., & Plaut, D. C. (2016). Disparate semantic ambiguity effects from semantic processing dynamics rather than qualitative task differences. *Language, Cognition and Neuroscience*, *31*(7), 940–966. https://doi.org/10.1080/23273798.2016.1171366

Armstrong, B. C., Tokowicz, N., & Plaut, D. C. (2012). Edom: Norming software and relative meaning frequencies for 544 english homonyms. *Behavior Research Methods*, *44*(4), 1015–1027. https://doi.org/10.3758/s13428-012-0199-8

Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature*, *533*(7604), 452–454. https://doi.org/10.1038/533452a

Beretta, A., Fiorentino, R., & Poeppel, D. (2005). The effects of homonymy and polysemy on lexical access: An meg study. *Cognitive Brain Research*, *24*(1), 57–65. https://doi.org/10.1016/j.cogbrainres.2004.12.006

Brysbaert, M., Mandera, P., & Keuleers, E. (2018). The word frequency effect in word processing: An updated review. *Current Directions in Psychological Science*, *27*(1), 45–50. https://doi.org/10.1177/0963721417727521

Brysbaert, M., & New, B. (2009). Moving beyond kučera and francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for american english. *Behavior Research Methods*, *41*(4), 977–990. https://doi.org/10.3758/BRM.41.4.977

Clark, H. H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior*, *12*(4), 335–359. https://doi.org/10.1016/S0022-5371(73)80014-3

Forster, K. I. (2000). The potential for experimenter bias effects in word recognition experiments. *Memory & Cognition*, *28*(7), 1109–1115. https://doi.org/10.3758/BF03211812

Frost, R., Armstrong, B. C., & Christiansen, M. H. (2019). Statistical learning research: A critical review and possible new directions. *Psychological Bulletin*, *145*(12), 1128.

Gernsbacher, M. A. (1984). Resolving 20 years of inconsistent interactions between lexical familiarity and orthography, concreteness, and polysemy. *Journal of experimental psychology: General*, *113*(2), 256.

Jager, B., & Cleland, A. (2015). Connecting the research fields of lexical ambiguity and figures of speech: Polysemy effects for conventional metaphors and metonyms. *Ment. Lexicon*, *10*(1), 133–151. https://doi.org/10.1075/ml.10.1.05jag

Li, J., & Joanisse, M. F. (2021). Word senses as clusters of meaning modulations: A computational model of polysemy. *Cognitive Science*, *45*(4), e12955. https://doi.org/10.1111/cogs.12955

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716. https://doi.org/10.1126/science.aac4716

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., ... Moher, D. (2021). The prisma 2020 statement: An updated guideline for reporting systematic reviews. *Systematic Reviews*, *10*(1), 89. https://doi.org/10.1186/s13643-021-01626-4

Pashler, H., & Wagenmakers, E.-.-J. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science*, *7*(6), 528–530. https://doi.org/10.1177/1745691612465253

Rodd, J., Gaskell, G., & Marslen-Wilson, W. (2002). Making sense of semantic ambiguity: Semantic competition in lexical access. *Journal of Memory and Language*, *46*(2), 245–266. https://doi.org/10.1006/jmla.2001.2810

Share, D. L. (2008). On the anglocentricities of current reading research and practice: The perils of overreliance on an "outlier" orthography. *Psychological Bulletin*, *134*(4), 584–615. https://doi.org/10.1037/0033-2909.134.4.584

Trott, S. (2024). Can large language models help augment english psycholinguistic datasets? *Behav-*

ior Research Methods. https://doi.org/10.3758/s13428-024-02337-z

Veritas Health Innovation. (2023). *Covidence systematic review software*. Melbourne, Australia. https://www.covidence.org

Wordsmyth. (2024). *Wordsmyth advanced dictionary*. Retrieved 2024, from https://www.wordsmyth.net/

Yarkoni, T. (2022). The generalizability crisis. *Behavioral and Brain Sciences*, *45*, e1. https://doi.org/10.1017/S0140525X20001685