Words as Gatekeepers: Measuring Discipline-specific Terms and Meanings in Scholarly Publications

Li Lucy[1,2] **Jesse Dodge**[1] **David Bamman**[2] **Katherine A. Keith**[1,3]
[1]Allen Institute for Artificial Intelligence
[2]University of California, Berkeley
[3]Williams College
{lucy3_li, dbamman}@berkeley.edu
jessed@allenai.org kak5@williams.edu

Abstract

Scholarly text is often laden with jargon, or specialized language that can facilitate efficient in-group communication within fields but hinder understanding for out-groups. In this work, we develop and validate an interpretable approach for measuring *scholarly jargon* from text. Expanding the scope of prior work which focuses on word types, we use word sense induction to also identify words that are widespread but overloaded with different meanings across fields. We then estimate the prevalence of these discipline-specific words and senses across hundreds of subfields, and show that word senses provide a complementary, yet unique view of jargon alongside word types. We demonstrate the utility of our metrics for science of science and computational sociolinguistics by highlighting two key social implications. First, though most fields reduce their use of jargon when writing for general-purpose venues, and some fields (e.g., biological sciences) do so less than others. Second, the direction of correlation between jargon and citation rates varies among fields, but jargon is nearly always negatively correlated with interdisciplinary impact. Broadly, our findings suggest that though multidisciplinary venues intend to cater to more general audiences, some fields' writing norms may act as barriers rather than bridges, and thus impede the dispersion of scholarly ideas.

1 Introduction

Specialized terminology, or jargon, naturally evolves in communities as members communicate to convey meaning succinctly. It is especially prevalent in scholarly writing, where researchers use a rich repertoire of lexical choices. However, niche vocabularies can become a barrier between fields (Vilhena et al., 2014; Martínez and Mammola, 2021; Freeling et al., 2019), and between scientists and the general public (Liu et al., 2022; August et al., 2020a; Cervetti et al., 2015; Freeling et al., 2021). Identifying scholarly jargon is an initial step for designing resources and tools that can increase the readability and reach of science (August et al., 2022a; Plavén-Sigray et al., 2017; Rakedzon et al., 2017).

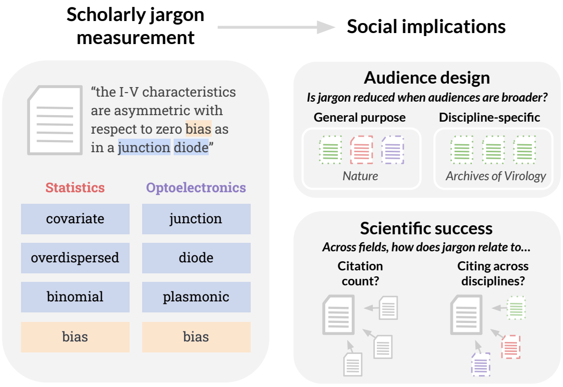

Scholarly jargon measurement
Social implications
"the I-V characteristics are asymmetric with respect to zero bias as in a junction diode"
Statistics
Optoelectronics
covariate
junction
overdispersed
diode
binomial
plasmonic
bias
bias
Audience design
Is jargon reduced when audiences are broader?
General purpose
Discipline-specific
Nature
Archives of Virology
Scientific success
Across fields, how does jargon relate to...
Citation count?
Citing across disciplines?

Figure 1: In this paper, we measure scholarly jargon, which consists of discipline-specific word types and senses. We further illustrate two social implications of jargon: whether its rate differs between broad and narrow audiences (right; top), and how it relates to citation-based success (right; bottom). The example abstract excerpt on the left is from Satishkumar et al. (2000).

Research on scholarly language typically focuses on the relative prevalence of words (McKeown et al., 2016; Prabhakaran et al., 2016; Sim et al., 2012; Rakedzon et al., 2017). However, the same word can be overloaded with multiple meanings, such as *bias* referring to electric currents or statistical misestimation (Figure 1). We use BERT-based word sense induction to disentangle these, and demonstrate the utility of including both word types and senses in our operationalization of *scholarly jargon*. We measure jargon in English abstracts across three hundred fields of study, drawn from over 12 million scholarly abstracts and one of the largest datasets of scholarly documents: the Semantic Scholar Open Research Corpus (S2ORC) (Lo et al., 2020).

Our findings are valuable for several groups that

Findings of the Association for Computational Linguistics: ACL 2023, pages 6929–6947
July 9-14, 2023 ©2023 Association for Computational Linguistics

partake in science: readers, authors, and science of science researchers.

Due to scholarly language's gatekeeping effect, natural language processing (NLP) researchers have developed tools to support **readers**, such as methods for simplifying or defining terminology (Kim et al., 2016; Vadapalli et al., 2018; August et al., 2022a; Head et al., 2021; August et al., 2022b; Murthy et al., 2022). When deciding what constitutes jargon, studies may rely on vocabulary lists based on word frequency, often collapsing all of science into one homogeneous language variety (August et al., 2020b; Rakedzon et al., 2017; Plavén-Sigray et al., 2017). Our approach identifies language associated with individual subfields and proposes a bottom-up, data-driven process for creating these vocabularies (§3 and 4).

Second, measuring levels of discipline-specific language in abstracts can inform **authors** who wish to communicate to a wider audience or enter a new field. We show that while some subfields tend to use highly specialized word types, others use highly specialized senses (§5). In addition, we provide evidence for audience design in scholarly discourse (§6.1), following a sociolinguistic framework that describes how speakers accommodate language to the scope of their audience (Bell, 1984).

Finally, our language-centered approach contrasts the typical paradigm in **science of science** research, where citation behavior often defines relationships among articles, venues, and fields (e.g. Boyack et al., 2005; Rosvall and Bergstrom, 2008; Peng et al., 2021). Citation count is a common measurement of "success", and the mechanisms behind it form a core research area (Wang and Barabási, 2021; Foster et al., 2015; Fortunato et al., 2018). On the other hand, interdisciplinarity is increasingly valued, but does not always lead to short-term citation gains (Van Noorden, 2015; Larivière and Gingras, 2010; Okamura, 2019; Chen et al., 2022). We run regression analyses to examine the relationship between discipline-specific senses and types and these two distinct measures of success (§6.2).

To summarize, we contribute the following (Figure 1):

- **Methods.** We propose a new measure of scholarly jargon to identify discipline-specific word types and senses (§3). We validate our approach for measuring senses by showing it recalls more overloaded words in Wiktionary compared to word types alone (Figure 3).

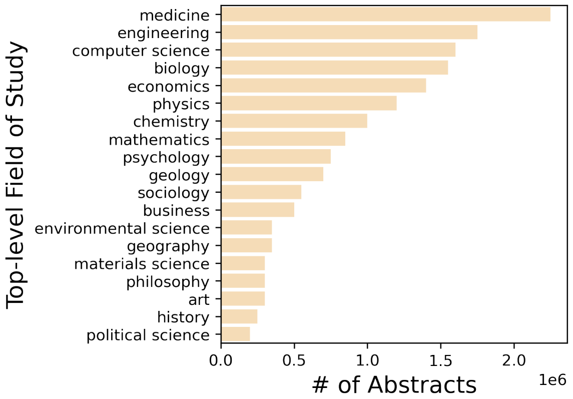


Figure 2: The number of abstracts in each top-level scholarly field in CONTEMPORARY S2ORC, sorted by size. In total, there are 12.0 million abstracts.

- **Social implications.** We illustrate the utility of our jargon measurements for computational social science by analyzing audience design and articles' success (§6). Though multidisciplinary venues may intend to be general-purpose, more dominant fields in these venues reduce jargon less so than others (Figure 5). Since jargon nearly always has a negative relationship with interdisciplinary impact (Table 4), our findings encourage the reconsideration of existing scholarly writing norms.

We hope our measure of scholarly jargon can help researchers quantify language barriers in science and their implications. Our code and scored lists of jargon for each subfield can be found at https://github.com/lucy3/words_as_gatekeepers.

2 Data

Our work involves several datasets: scholarly abstracts, Wikipedia, and Wiktionary. We use abstracts to calculate the association of words with disciplines and Wikipedia to supplement our calculation of background word probabilities. Later, in §4, we introduce and describe how we use Wiktionary to validate our approach.

2.1 Contemporary S2ORC

Our dataset of academic articles, CONTEMPORARY S2ORC, contains 12.0 million abstracts and 2.0 billion words[1] that span a mix of scholarly fields (Figure 2, Appendix A).

[1] We define a word as a non-numeric, non-punctuation token outputted by Huggingface transformer's whole-word BasicTokenizer.

To create CONTEMPORARY S2ORC, we draw from the July 2020 release of S2ORC (Lo et al., 2020). S2ORC is a general purpose corpus that contains metadata for 136 million scholarly articles, including 380.5 million citation links (Lo et al., 2020). These articles originate from Semantic Scholar, which obtains data directly from publishers, the Microsoft Academic Graph (MAG), arXiv, PubMed, and the open internet. Metadata, such as titles, authors, publication years, journals/venues, and abstracts are extracted from PDFs and LaTeX sources or provided by the publisher. Though extensive, S2ORC contains some amount of noisy or missing metadata. We remove non-English articles and those with missing metadata, consolidate journals and venues into a single venue label, and limit the dataset to the years 2000-2019 (Appendix B).

S2ORC links articles to paper IDs in the Microsoft Academic Graph (MAG) (Sinha et al., 2015; Wang et al., 2019), so we match S2ORC abstracts to MAG fields of study (FOS). S2ORC originally contains top-level MAG FOS (level 0), e.g. *biology*, but we also join abstracts with second level MAG FOS (level 1), e.g. *immunology*, for more granularity.[2] In this present paper, we refer to level 0 FOS as *fields*, and level 1 FOS as *subfields*. We take an approximately uniform sample of 50k articles per subfield, resulting in a total of 293 subfields that fall under 19 fields (Appendix A).

2.2 Wikipedia

We include Wikipedia article content to counterbalance CONTEMPORARY S2ORC's STEM-heavy focus for our estimation of words' typical prevalence. Wikipedia is a popular information-gathering resource (Reagle and Koerner, 2020), and we use an Oct 1, 2022 dump of its articles. It offers complementary topical coverage that is collectively curated and driven by public interest, and includes biographies, culture, and arts (Mesgari et al., 2015). We remove Wiki formatting using Attardi (2015)'s text extractor, and discard all lines, or paragraphs, that are less than 10 white-spaced tokens long. We sample twice as many Wikipedia paragraphs as the number of CONTEMPORARY S2ORC abstracts, so that each is similar in size despite differences in document length. In total our Wikipedia dataset, WIKISAMPLE, contains 24.0 million paragraphs and 1.5 billion tokens.

3 Methods

Language differences among subsets of data can be measured by a variety of approaches, from geometric to information theoretic (Ramesh Kashyap et al., 2021; Vilhena et al., 2014; Aharoni and Goldberg, 2020). We calculate the association of a word's type or sense to subfields using normalized pointwise mutual information (NPMI). We choose NPMI over similar metrics (e.g. tf-idf, divergence, z-score) because of the nature of language difference it emphasizes: higher NPMI scores reflect language that is not only commonly used in a community, but also highly specific to it (Lucy and Bamman, 2021; Gardner et al., 2021). NPMI offers an interpretable metric of association, where a score of 1 indicates perfect association, 0 indicates independence, and -1 indicates no association. We follow Lucy and Bamman (2021)'s framework of calculating NPMI separately for word types and senses, which they originally used to identify community-specific language on social media. We update their approach with a more recent word sense induction (WSI) method, and use a different interpretation of type and sense NPMI scores.

3.1 Discipline-specific words

We calculate NPMI for word types, or *type NPMI*, as the following measure::

$$\mathcal{T}_f(t) = \frac{\log\left(P(t \mid f)/P(t)\right)}{-\log P(t,f)}. \quad (1)$$

Here, $P(t \mid f)$ is the probability of a word t occurring given a set of abstracts f in a field, $P(t,f)$ is their joint probability, and $P(t)$ is the probability of the word overall (Lucy and Bamman, 2021; Zhang et al., 2017). "Overall" refers to the combined background dataset of CONTEMPORARY S2ORC and WIKISAMPLE. We only calculate $\mathcal{T}_f(t)$ for words that appear at least 20 times in each field.[3]

As illustrative examples, Table 1 shows words with the highest $\mathcal{T}_f(t)$ in several fields.

3.2 Discipline-specific senses

Widely disseminated words can be overloaded with domain-specific meanings or use. For example,

[2] A secondary level FOS may fall under multiple top-level FOS, some articles are labeled with multiple FOS at the same level, and some articles marked with top-level FOS do not indicate a secondary level FOS.

[3] As we will describe in §3.2, the sense NPMI pipeline operates on lemmas, not words. Standard lemmatizers may not be suitable for rarer words in science, so to make our type and sense metrics comparable, we only lemmatize the set of widely-used words that are shared by both pipelines.

NLP		Chemical Engineering		Immunology		Communication		International Trade		Epistemology	
word	$\mathcal{T}_f(t)$	word	$\mathcal{T}_f(t)$	word	$\mathcal{T}_f(t)$	word	$\mathcal{T}_f(t)$	word	$\mathcal{T}_f(t)$	word	$\mathcal{T}_f(t)$
nlp	0.412	rgo	0.334	treg	0.346	saccade	0.354	wto	0.453	epistemic	0.356
corpora	0.404	mesoporous	0.328	cd4	0.341	saccades	0.345	trade	0.438	epistemology	0.350
treebank	0.401	nanosheets	0.327	immune	0.3388	stimuli	0.333	fdi	0.401	epistemological	0.342
disambiguation	0.396	nanocomposite	0.325	il	0.336	stimulus	0.331	ftas	0.396	husserl	0.332
corpus	0.393	nanocomposites	0.324	th2	0.335	cues	0.327	antidumping	0.396	kant	0.329

Table 1: Top five words that are highly specialized to different disciplines. These have the highest type NPMI ($\mathcal{T}_f(t)$) scores in their respective subfields. As examples, *treg* in immunology stands for "regulatory T cells", and *antidumping* in international trade places high taxes on imports.

bias could refer to a type of voltage applied to an electronic system, social prejudice, or statistical misestimation. Thus, we include word senses as a complement to word types for characterizing domain-specific language. We use *senses* to refer to different meanings or uses of the same word induced by word sense induction (WSI).

3.2.1 Word sense induction

To partition occurrences of words into senses, we adapt Eyal et al. (2022)'s WSI pipeline with minimal modifications. WSI is an unsupervised task where occurrences of words are split into senses. Eyal et al. (2022)'s approach is designed for large-scale datasets, where a sample of a target word's occurrences is used to induce senses, and remaining occurrences are then assigned to them. To induce senses, a masked language model predicts the top s substitutes of each occurrence of a target word. Then, a network is created for each target word, where nodes are substitutes and edges are their co-occurrence. Louvain community detection is then applied to determine senses, or sets of substitutes (Blondel et al., 2008). For example, in the network for *bass*, the substitutes for its sense as a type of fish are likely not predicted at the same time as substitutes for its musical sense, so each set would represent separate senses.

We carry out this WSI pipeline on a case-insensitive target vocabulary of 6,497 "widely used" words: those that appear in the top 98th percentile by frequency and in at least 50% of venues, not including stopwords and words split into wordpieces.[4] We lemmatize and lowercase target words and substitutes, following Eyal et al. (2022)'s implementation, because otherwise the most common substitutes representing a sense may be different lemmas of the same word. This processing step reduces the target vocabulary into 4,407 lemmas.

We sample 1000 instances of each vocabulary lemma, and use ScholarBERT to predict each instance's top $s = 5$ substitutes (Hong et al., 2022).[5] We truncate each abstract to this model's maximum input length. We follow Eyal et al. (2022)'s heuristics for determining sets of substitutes that are big enough to recognize as senses: each set needs to have at least two substitutes, and the second most frequent substitute needs to appear at least 10 times across the target word's sample. If no sets are big enough, we add a fallback case, where we place all occurrences of a word to a single sense.

Eyal et al. (2022) assigns additional occurrences of the target word to induced senses based on Jaccard similarity. We also add a fallback case here: if the overlap of a remaining occurrence's substitutes with all senses is zero, we assign that occurrence to an extra sense representing previously unseen senses.

3.2.2 Sense NPMI

Once each occurrence of a widely-used word is labeled with a sense, their frequencies can be used to calculate sense NPMI. Sense NPMI uses the same formula as type NPMI, except it is calculated at the sense-level rather than the word-level (Lucy and Bamman, 2021). That is, counts of a word t are replace with counts of its ith sense, t_i:

$$\mathcal{S}_f(t_i) = \frac{\log\left(P(t_i \mid f)/P(t_i)\right)}{-\log P(t_i, f)}. \qquad (2)$$

4 Validation

4.1 Wiktionary

We perform in-domain validation of the unsupervised sense pipeline using Wiktionary. Words

[4] We avoid wordpieces since Eyal et al. (2022)'s pipeline predicts substitutes at the token-level.

[5] We pick ScholarBERT over similar transformer models trained on scholarly language (e.g. SciBERT), because it is trained on a wider breadth of disciplines, splits fewer potential vocab words into wordpieces (15 versus 193), and uses RoBERTa-style training (Hong et al., 2022; Beltagy et al., 2019).

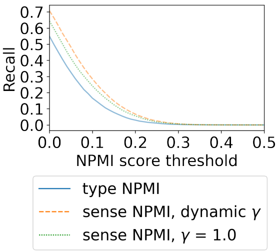
marked as associated with a subfield in this online dictionary should also be highly scored by our metrics. Wiktionary is collaboratively maintained and includes common words listed with definitions that may be labeled as having “restricted usage” to a topic or context. For example, the word *ensemble* has the labels *machine learning*, *fashion*, and *music* (Appendix C). We map Wiktionary labels in English definitions of target words using exact string matching to fields and subfields. If an NPMI score threshold were used to determine whether a token should be considered discipline-specific or not, we expect sense NPMI to recall more words labeled by Wikitionary than type NPMI does. We do not calculate precision, because Wiktionary is not necessarily comprehensive for all subfields.

We obtain Wiktionary entries for 94.94% of the common, widely used words that were inputs in the WSI pipeline. We filter out words where all definitions are labeled with only one field, and allow subfields to inherit the words labeled with their parent field. In total, we have 11,548 vocabulary word and subfield pairs to recall across 83 subfields. Since recall is calculated at the word-level and sense NPMI is at the sense-level, we use a word t's most frequent sense t_i's $\mathcal{S}_f(t_i)$ in a subfield to represent word-level sense NPMI $\mathcal{S}_f(t)$.

In Eyal et al. (2022)'s WSI pipeline, the resolution parameter γ in Louvain community detection calibrates the number of senses induced per word. Increasing resolution leads to more fine-grained word senses and higher recall, but potentially spurious senses (Figure 3). Rather than using Eyal et al. (2022)'s default resolution value of 1, we use a dynamic formula for resolution (Newman, 2016):

$$\gamma = \frac{\omega_{in} - \omega_{out}}{\log \omega_{in} - \log \omega_{out}}, \tag{3}$$

where ω_{in} is the probability of an edge between two nodes in the same community, and ω_{out} is the probability of an edge between two nodes in different communities. Intuitively, nodes within communities should be more connected than nodes between them. We follow Newman (2016)'s algorithm, initializing $\gamma = 1$ and iterating for each target lemma at most 10 times. In each iteration, we run Louvain community detection and recalculate γ using the edge probabilities in the current clustering. We stop early if γ converges within 0.01 of its previous value.



NPMI metric	AUC, recall
$\mathcal{S}_f(t)$, $\gamma = 0.5$	0.0550
$\mathcal{S}_f(t)$, $\gamma = 1.0$	0.0583
$\mathcal{S}_f(t)$, $\gamma = 1.5$	0.0631
$\mathcal{S}_f(t)$, $\gamma = 2.0$	0.0670
$\mathcal{S}_f(t)$, $\gamma = 2.5$	0.0697
$\mathcal{S}_f(t)$, dynamic γ	0.0675
$\mathcal{T}_f(t)$ baseline	0.0434

Figure 3: Recall and area under the curve (AUC) of 11,548 Wiktionary words with discipline-specific definitions. Sense NPMI with dynamic resolution (γ) recalls more semantically overloaded words than type NPMI at the same score threshold.

Sense NPMI with dynamic resolution recalls more discipline-specific Wiktionary words than type NPMI at the same score cutoff (Figure 3). In addition, the sense NPMI of a word in a subfield labeled by Wiktionary is higher than the score of the same word in a random field (paired t-test, $p < 0.001$, Appendix C). Thus, Wiktionary-based validation shows that our unsupervised approach is able to measure discipline-specific senses, and in all downstream analyses, we use the dynamically defined γ for WSI.

4.2 Examples and interpretation

Examples of semantically overloaded words between fields can also lend face validity to our results (Table 2). Returning to the example introduced at the beginning, *bias* is indeed very overloaded. It has distinct senses with high NPMI (> 0.2) across multiple fields, including statistics (*skew*),[6] optoelectronics (*charge*), cognitive psychology (*preference*), and climatology (*error*). These examples suggest that future work could examine how our approach could provide potential candidates for updating dictionaries or glossaries when new senses are introduced.

Table 3 shows examples of words whose scores increase from type NPMI to sense NPMI despite having counts split across senses. Lucy and Bamman (2021) interpret sense and type NPMI similarly in their downstream analyses, based on the magnitude of their values, but this does not account for how type and sense NPMI scores are related. In the boundary case where a word t only has a single sense t_0, $\mathcal{S}_f(t_0) = \mathcal{T}_f(t)$. This leads to a strong correlation between the two metrics, especially when a sense scored as highly associated with a field is also the dominant sense of that word in general. Thus, to narrow in on what

[6] Word in parentheses is the top predicted substitute for that subfield's sense for *bias*.

word t	sense t_1 FOS a	$\mathcal{S}_a(t_1)$	top substitutes	sense t_2 FOS b	$\mathcal{S}_b(t_2)$	top substitutes
kernel	Operating system	0.321	block, personal, ghost, every, pure	Agronomy	0.272	grain, palm, body, gross, cell
performance	Chromatography	0.266	perform, play, timing, temperature, contribute	Industrial organization	0.234	success, record, position, accomplishment, hand
network	Computer network	0.327	graph, net, regular, key, filter	Telecommunications	0.259	connection, channel, link, connectivity, association
root	Dentistry	0.413	crown, arch, tooth, long, tissue	Horticulture	0.330	plant, tree, branch, part, stem
power	Electrical engineering	0.329	energy, electricity, load, fuel, lit	Combinatorics	0.193	value, order, term, sum, degree

Table 2: Hand-selected words that are common across fields, but have different uses or meanings. The senses shown for each word are the two with the highest sense NPMI scores for that word across fields. Each sense is represented by the five most common substitutes suggested by ScholarBERT for instances in that sense.

Pure mathematics				Monetary economics				Computer security				Stereochemistry			
word	Δ	$\mathcal{S}_f(t)$	$\mathcal{T}_f(t)$	word	Δ	$\mathcal{S}_f(t)$	$\mathcal{T}_f(t)$	word	Δ	$\mathcal{S}_f(t)$	$\mathcal{T}_f(t)$	word	Δ	$\mathcal{S}_f(t)$	$\mathcal{T}_f(t)$
power	0.202	0.186	-0.016	movement	0.218	0.266	0.048	primitive	0.162	0.221	0.058	attack	0.228	0.184	-0.044
pole	0.194	0.207	0.013	liquid	0.195	0.196	0.002	host	0.151	0.205	0.054	title	0.216	0.264	0.048
union	0.193	0.141	-0.051	interest	0.182	0.382	0.200	elasticity	0.148	0.158	0.010	km	0.212	0.175	-0.037
surface	0.193	0.260	0.068	turbulence	0.176	0.155	-0.021	hole	0.147	0.134	-0.013	framework	0.205	0.215	0.010
origin	0.193	0.188	-0.005	provider	0.176	0.121	-0.055	key	0.142	0.320	0.179	solve	0.202	0.165	-0.037

Table 3: Top five words that have senses associated with each subfield ($\mathcal{S}_f(t) > 0.1$), ordered by the difference Δ between word-level sense and type NPMI. These are words that are highly specific to subfields based on their sense, rather than their type. As examples, monetary economics uses *liquid* to describe valuables that can be easily converted to cash, and stereochemistry uses *attack* to refer to the addition of atoms or molecules during chemical reactions.

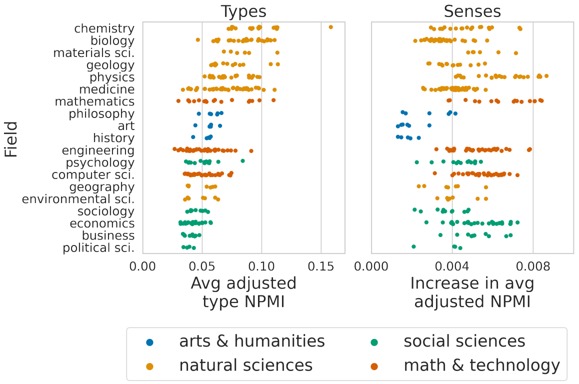


Figure 4: The left subplot shows the distinctiveness of subfields' word types, while the right shows the increase in distinctiveness when we take the max of words' type and sense NPMI instead of only their type NPMI. Each point is a subfield, such as *organic chemistry* in *chemistry*, and fields are colored using larger disciplinary categories for interpretation clarity.

we gain from WSI, we examine not only senses that are highly associated with a field, but have sense NPMI scores higher than their words' type NPMI scores (Table 3). Therefore, we count a token with a labeled sense as a *discipline-specific sense* if $\mathcal{S}_f(t_i) > \mathcal{T}_f(t)$ and $\mathcal{S}_f(t_i) > c$ for a subfield f and some cutoff c. Otherwise, the token is a *discipline-specific type* if $\mathcal{T}_f(t) > c$.

5 Language norms across fields

The linguistic insularity of science varies across fields. For example, Vilhena et al. (2014) found that phrase-level jargon separates biological sciences more so than behavioral and social sciences. We perform a similar analysis with the novel addition of word senses.

To summarize the distinctiveness of word types in a field, we calculate the mean type NPMI score of unique words in a field. Before taking the mean, however, we adjust scores by zeroing negative values, since we are more interested in words associated with a field rather than those that are not. This zeroing practice is typically used in studies where PMI measures word relatedness (Levy et al., 2015; Dagan et al., 1993; Bullinaria and Levy, 2007).

Like Vilhena et al. (2014), we also find that the biological sciences have very distinctive word types (Figure 4). However, there is a considerable amount of overlap in word type distinctiveness across fields. Similar to how natural sciences name molecules and chemicals, the arts and humanities name canons of writers, philosophers, and artists.

We also examine what fields gain the most in NPMI scores when common words are broken into their senses. We recalculate subfields' average adjusted NPMI, but use $\max(\mathcal{T}_f(t), \mathcal{S}_f(t))$ instead of $\mathcal{T}_f(t)$ for words that have induced senses. Based on their relative increases in average adjusted NPMI, subfields in math/technology, physics, and economics often use common words in specialized contexts (Table 3, Figure 4). There is no significant

Pearson correlation between the distinctiveness of subfields' word types and that of their senses. Thus, word senses provide a very different perspective on language norms and suggests an additional route through which gatekeeping may occur.

6 Social implications

In this section, we examine two social implications of our metrics: audience design and scholarly success. We limit these experiments to articles in CONTEMPORARY S2ORC that are published among 11,047 venues in the top 95th percentile by abstract count (at least 800 each in S2ORC), to ensure solid estimation of venue-level information, such as their disciplinary focus and average citations per article.

6.1 Audience design

Audience design is a well-studied sociolinguistic phenomenon where a speaker's language style varies across audiences (Bell, 1984, 2002; Ndubuisi-Obi et al., 2019; Androutsopoulos, 2014). For example, on Twitter, when writers target smaller or more geographically proximate audiences, their use of nonstandard language increases (Pavalanathan and Eisenstein, 2015). Here, we examine this type of language accommodation at the level of subfields, as our data does not contain unique author identifiers that would allow measurements of author-level variation. We hypothesize that for abstracts within the same subfield, ones published for broader audiences (general-purpose venues) use less scholarly jargon than those published in narrower, discipline-focused venues.

To address this hypothesis, we first collect sets of 6 general-purpose and 2464 discipline-specific venues. We use general-purpose venues that appear in both our dataset and Wikipedia's list of general and multidisciplinary journals:[7] *Nature*, *Nature Communications*, *PLOS One*, *Science*, *Science Advances*, and *Scientific Reports*. Discipline-focused venues are those where 80% of articles fall under a single subfield or its name contains the subfield, e.g. *Agronomy Journal*.[8] Among these two venue sets, we examine abstracts labeled with only one subfield.

We then calculate the fraction of jargon over all words in each abstract, by counting tokens t that are either discipline-specific senses or types, where $c = 0.1$. In other words, we count t if $\max(\mathcal{T}_f(t), \mathcal{S}_f(t)) > 0.1$ in the abstract's subfield f. We find that most fields adjust their rate of jargon based on audience, though fields such as medicine and physics are notable exceptions (Figure 5). One explanation for this exceptional behavior is that general-purpose venues have a history of being led and dominated by biological sciences, and in some, by physical sciences as well (de Carli and Pereira, 2017; Koopman, 2011; Varmus et al., 2000). Thus, jargon-laden fields further from these areas adjust their writing the most when publishing in these venues.

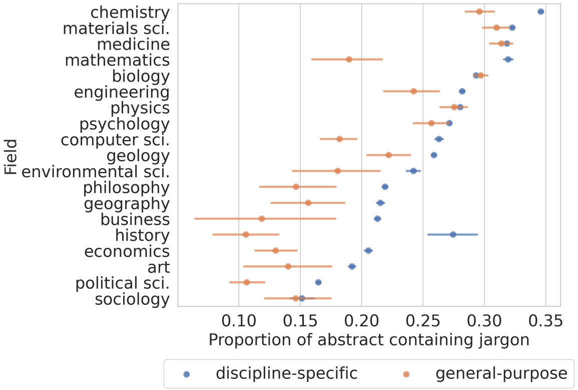


Figure 5: Abstracts in the same field typically contain less jargon when they appear general-purpose venues (e.g. *Nature*) than when they are in discipline-focused ones (e.g. *Genetics*). Fields are ordered by monotonically decreasing averages, and error bars are 95% CI.

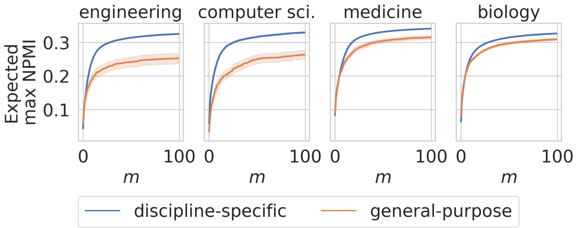


Figure 6: The expected maximum type and sense NPMI of tokens in abstracts, where m on the x-axis indicates token position and shaded areas are 95% CI. Computer science and engineering have wider gaps in scholarly jargon use between general-purpose and discipline-focused venues than medicine and biology do.

A limitation of this approach for quantifying the amount of jargon in an abstract is that it relies on choosing c. We also obtain similar results with $c = 0.2$ and justify our choice of c in Appendix D.1. An alternative perspective mimics how soon a reader may encounter highly special-

[7] https://en.wikipedia.org/wiki/List_of_scholarly_journals.

[8] There is substantial overlap between these two groups, where 80% of venues dominated by one subfield also mention the subfield in its name.

ized language in an abstract. In this approach, we calculate the maximum over an abstract's type or sense NPMI scores within the first m tokens of the abstract. These results provide another view of our previous finding: fields such as computer science and engineering adjust their content for general-purpose venues more so than those in the biological sciences (Figure 6). This indicates that though most "general-purpose" venues intend to be for all of science,[9] some fields are expected to adapt their language more so than others.

Among in-group members, the use of specialized vocabulary can signal legitimacy and expertise (Agha, 2005; Labov, 1973). Thus, there may be competing incentives influencing authors' writing. In the next section, we further investigate the relationship between jargon and two incentives in science: citation count and a metric of interdisciplinary impact.

6.2 Scholarly success

We hypothesize that jargon plays different roles in the success of an article depending on how "success" is defined. In particular, since jargon gatekeeps outsiders from a discipline, we expect it to negatively affect interdisciplinary impact. To test this hypothesis, we run two sets of regressions to measure the relationship between abstracts' use of jargon and citation behavior within five years after publication. The first set of regressions predicts short-term citation counts, while the second predicts interdisciplinary impact. We run separate regression models for each field to compare heterogeneity across fields. Each unit of analysis is an abstract published in 2000-2014 labeled with only one or two subfields.

Two key independent variables are the fractions of discipline-specific words and senses in an abstract, with $c = 0.1$. For abstracts that have two subfields in the same analyzed parent field, we sum their type and sense jargon counts. Additional independent variables include time (three evenly-sized time bins within 2000-2014), length of abstract in tokens, number of authors, number of references in the article, number of subfields (one or two), and the venue's average citations per article.

	Citation count			**Interdisciplinary impact (DIV)**		
Field	types	senses	# obv.	types	senses	# obv.
Medicine	-0.15***	0.60***	1,137,923	-0.10***	-0.05***	589,641
Engineering	0.07	0.64***	786,559	-0.09***	-0.15***	199,790
Comp. sci.	-0.87***	0.71***	556,330	-0.12***	-0.11***	196,234
Biology	-0.12***	0.52***	824,768	-0.80***	-0.03***	481,103
Economics	0.15	1.23***	454,215	-0.11***	0.00	123,476
Physics	0.47***	-1.04***	648,729	-0.16***	-0.10***	203,009
Chemistry	-1.36***	-2.32***	613,535	-0.10***	-0.08***	187,621
Mathematics	1.22***	1.40***	363,369	-0.15***	-0.11***	128,482
Psychology	0.34***	3.68***	261,102	-0.11***	-0.06***	133,319
Geology	-0.42***	0.83***	343,250	-0.13***	-0.13***	138,308
Sociology	1.18***	2.24***	149,484	-0.08***	0.01	56,088
Business	0.30**	2.71***	160,536	-0.11***	-0.04***	39,602
Environ. sci.	-1.22***	-2.20***	137,862	-0.12***	-0.05***	49,199
Geography	0.17	0.37	127,561	-0.10***	-0.04***	51,408
Material sci.	-1.73***	1.42***	149,602	-0.14***	-0.09***	45,445
Philosophy	-0.92***	2.16***	68,512	-0.03***	0.06***	10,559
Art	-1.75***	-2.30	68,220	-0.04***	0.03	5,826
History	-0.27	10.94***	47,910	-0.50***	0.05	6,513
Political sci.	2.27***	2.86***	44,994	-0.04**	0.03	8,486

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$ with Bonferroni correction.

Table 4: The columns *type* and *sense* show regression coefficients for the fractions of discipline-specific words or senses in abstracts. The dependent variables are citation count and interdisciplinary impact (DIV). Significantly negative coefficients are highlighted, and # obv. is the number of observations. Since dependent variables and their expected values differ across regressions, the magnitude of coefficients are not comparable.

Citation count is an over-dispersed count variable, so we run a negative binomial regression to predict this outcome (Hilbe, 2011). In some cases, jargon use has a significant positive relationship with citations, but the direction of this relationship differs across fields (Table 4, Appendix D.2).

Alternatively, interdisciplinary impact considers the subfield composition of articles citing a target abstract. We use Leydesdorff et al. (2019)'s established formula, which they call DIV:

$$\mathrm{DIV}(\mathcal{C}) = \frac{n}{N}(1 - \mathrm{Gini}) \sum_{i,j \in \mathcal{C}, i \neq j} \frac{d_{ij}}{n(n-1)}, \quad (4)$$

where $\mathcal{C}$ is the set of subfields citing the abstract, $n = |\mathcal{C}|$, N is the total number of subfields, and $d_{ij} = 1 - \cos(v_i, v_j)$, where v are subfields vectorized using overall cross-subfield citation counts (Appendix D.2). The first component measures the fraction of citing subfields, the second uses the Gini coefficient to calculate balance of citation counts among $\mathcal{C}$, and the third incorporates subfield similarity (Leydesdorff et al., 2019; Chen et al., 2022; Stirling, 1998). We run ordinary least squares regression on abstracts that are cited by at least two subfields, with *DIV* as the dependent variable. Discipline-specific words and senses have a negative relationship with *DIV* across fields that have highly distinctive language norms (Table 4).

Thus, though jargon has a varying relationship with citation counts, our regression results suggest

[9] For example, *Nature*'s "Aims and Scope": https://www.nature.com/nature/journal-information.

that it may generally impede the forging of interdisciplinary connections.

7 Related Work

Computational sociolinguistics often focuses on social media (Nguyen et al., 2016), with less attention on situation-dependent language varieties, or *registers*, in scholarly communities (Agha, 2005). Here, language differences can indicate different factions of authors and disciplinary approaches (Ngai et al., 2018; West and Portenoy, 2016; Sim et al., 2012). In addition to our present work, a few studies have examined word meaning or use, such as semantic influence or novelty (Soni et al., 2021, 2022) and semantic uncertainty (McMahan and Evans, 2018). Research on lexical ambiguity in science also appears in education, with an emphasis on how to improve the teaching of overloaded terminology (Ryan, 1985; Cervetti et al., 2015). Other NLP studies of science have predicted responses to articles (Yogatama et al., 2011), measured impact and innovation (Gerow et al., 2018; Hofstra et al., 2020; McKeown et al., 2016), and classified topics' rhetorical functions (Prabhakaran et al., 2016).

8 Conclusion

We use data-driven, interpretable methods to identify jargon, defined as discipline-specific word types and senses, across science at scale. By identifying senses, we are able to recall more words labeled as associated with a field in Wiktionary than with word types alone. We then map language norms across subfields, showing that fields with distinctive word types differ from those with distinctive word senses. Finally, we analyze implications of jargon use for communication with out-groups. We find that supposedly general-purpose venues have varying expectations around abstracts' use of jargon depending on the field, and jargon is negatively related to interdisciplinary impact. This suggests a potential opportunity for the reconsideration of abstract writing norms, especially for venues that intend to bridge disciplines.

9 Limitations

Below, we outline several limitations of our work.

Data coverage. Our claims are only valid for the datasets accessed in our study. We use the Microsoft Academic Graph (Sinha et al., 2015) and S2ORC, which is larger than other publicly-available scientific text corpora (Lo et al., 2020). However, these sources can differ from other collections of scientific text, because which journal/venues, sources, and resource types constitute "science" differs across academic literature search systems and databases (Gusenbauer and Haddaway, 2020; Ortega and Aguillo, 2014). In particular, since a substantial portion of S2ORC comes from scrapes of arXiv and PubMed, its coverage of computer science and medicine is better than that of other fields (Lo et al., 2020). Also, our coverage is limited to English articles. Past work has shown that citation-based metrics of impact favor articles written in English, and articles from non-English-speaking countries have different citation patterns compared to others (Liang et al., 2013; Liu et al., 2018; González-Alcaide et al., 2012). Finally, we recognize that MAG field of study labels are contestable and imperfect. For example, less than two-thirds of *ACL* articles are labeled as *natural language processing*, and the most popular subfield in *ICML* is *mathematics* rather than *machine learning*.

Token-level analyses. Another limitation of our study is that many scholarly terms are not single words or tokens, but rather phrases. Phrases are somewhat accounted for by measuring words' senses, since senses induced by language models reflect words' in-context use, including their use in discipline-specific phrases. For example, Table 3 shows that *title* has a sense specific to stereochemistry, and in abstracts, this word often occurs in the phrases *title reaction* or *title compound*. Phrases containing distinctive words are also somewhat accounted for by measuring individual words in the phrase. However, phrase-level measurements of jargon would likely still be useful for improving interpretability and downstream applications of our metrics, and so discipline-specific phrases are a promising avenue for future work.

Compute. Science of science is interdisciplinary and involves a range of organizations and institutions. Not all researchers will have easy access to the computational resources needed to replicate our study or apply our approach to data of the same scale. The most resource intensive step of our pipeline is when ScholarBERT predicts each instance of a vocabulary word's top 5 substitutes across CONTEMPORARY S2ORC and WIKISAMPLE. This took approximately 90 GPU hours split across Nvidia RTX A6000 and Quadro RTX 8000 GPUs. ScholarBERT itself is a 770M-parameter BERT model (Hong et al., 2022), and generally our compute infrastructure included machines with 64 to 128 cores and 512 to 1024 GB of RAM.

Social implications. In §6.2, we define "success" in two ways, both of which are based on citations. However, though citations are an important currency in science, they are imperfect signals of credit or impact. One article may cite another for reasons that span a range of significance, from brief mentions of related background to core motivation (Jurgens et al., 2018). In addition, associations between jargon use and scientific success may differ as success is redefined using indicators beyond citations. For example, success could be defined beyond scientific communities, such as findings that lead to societal change, products, and use (Bornmann, 2013). Finally, our study on the relationship between jargon and success is not causal, but associational and descriptive.

10 Ethical considerations

Data. With regards to data privacy, the dataset we use, S2ORC, is not anonymized, since entries for each article includes a list of author names. Even with the removal of author names, data can easily be linked to authors since abstracts are published online with attribution. We don't use author information in our research, and our outputs are aggregated over subsets of data. Still, we acknowledge that science of science research involving author information has the risk of judging research productivity and quality using metrics that may deemphasize some forms of contribution and labor, systemically disadvantaging some demographic groups. In addition, we did not receive the explicit consent of authors to use their content for our study, though the harms of this are minimized since the type of science we study is inherently a public-facing endeavor. S2ORC is released under a CC BY-NC 4.0 license, and its intended use is for NLP task development and science of science analysis. Any derivatives we produce share the same intended use and license.

"Jargon". In this paper, we use *jargon* to refer to sets of words that are specific to a discipline. *Jargon* can be a neutral term when referring to scientific or technical language, but has negative connotations of being incomprehensible or undesirable when used to refer to community vernacular or entire language varieties. Thus, care should be taken when deciding when and how to use *jargon* to refer to language.

11 Acknowledgements

We thank Misha Teplitskiy, Sandeep Soni, Isaac Bleaman, and Tal August for helpful conversations during the completion of this paper, and the Semantic Scholar team for their support in using and managing data. In addition, we thank our anonymous reviewers for their feedback. KK is grateful for the support of a Young Investigator Grant from the Allen Institute for Artificial Intelligence.

References

Asif Agha. 2005. *Registers of Language*, chapter 2. John Wiley & Sons, Ltd.

Roee Aharoni and Yoav Goldberg. 2020. Unsupervised domain clusters in pretrained language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7747–7763, Online. Association for Computational Linguistics.

Jannis Androutsopoulos. 2014. Languaging when contexts collapse: Audience design in social networking. *Discourse, Context & Media*, 4-5:62–73. Digital language practices in superdiversity.

Giuseppe Attardi. 2015. Wikiextractor. `https://github.com/attardi/wikiextractor`.

Tal August, Dallas Card, Gary Hsieh, Noah A. Smith, and Katharina Reinecke. 2020a. Explain like I am a scientist: The linguistic barriers of entry to r/science. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–12, New York, NY, USA. Association for Computing Machinery.

Tal August, Lauren Kim, Katharina Reinecke, and Noah A. Smith. 2020b. Writing strategies for science communication: Data and computational analysis. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5327–5344, Online. Association for Computational Linguistics.

Tal August, Katharina Reinecke, and Noah A. Smith. 2022a. Generating scientific definitions with controllable complexity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8298–8317, Dublin, Ireland. Association for Computational Linguistics.

Tal August, Lucy Lu Wang, Jonathan Bragg, Marti A. Hearst, Andrew Head, and Kyle Lo. 2022b. Paper plain: Making medical research papers approachable to healthcare consumers with natural language processing.

Allan Bell. 1984. Language style as audience design. *Language in Society*, 13(2):145–204.

Allan Bell. 2002. *Back in style: reworking audience design*, page 139–169. Cambridge University Press.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008.

Lutz Bornmann. 2013. What is societal impact of research and how can it be assessed? A literature survey. *Journal of the American Society for Information Science and Technology*, 64(2):217–233.

Kevin W Boyack, Richard Klavans, and Katy Börner. 2005. Mapping the backbone of science. *Scientometrics*, 64(3):351–374.

T. S. Breusch and A. R. Pagan. 1979. A simple test for heteroscedasticity and random coefficient variation. *Econometrica*, 47(5):1287–1294.

John A Bullinaria and Joseph P Levy. 2007. Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior Research Methods*, 39(3):510–526.

A Colin Cameron and Pravin K Trivedi. 2013. *Regression Analysis of Count Data*. Cambridge University Press.

Gina N. Cervetti, Elfrieda H. Hiebert, P. David Pearson, and Nicola A. McClung. 2015. Factors that influence the difficulty of science words. *Journal of Literacy Research*, 47(2):153–185.

Shiji Chen, Yanhui Song, Fei Shu, and Vincent Larivière. 2022. Interdisciplinarity and impact: The effects of the citation time window. *Scientometrics*, 127(5):2621–2642.

Ido Dagan, Shaul Marcus, and Shaul Markovitch. 1993. Contextual word similarity and estimation from sparse data. In *31st Annual Meeting of the Association for Computational Linguistics*, pages 164–171, Columbus, Ohio, USA. Association for Computational Linguistics.

Gabriel José de Carli and Tiago Campos Pereira. 2017. Multidisciplinarity: Widen discipline span of nature papers. *Nature*, 545(7654):289–289.

Matan Eyal, Shoval Sadde, Hillel Taub-Tabib, and Yoav Goldberg. 2022. Large scale substitution-based word sense induction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4738–4752, Dublin, Ireland. Association for Computational Linguistics.

Santo Fortunato, Carl T. Bergstrom, Katy Börner, James A. Evans, Dirk Helbing, Staša Milojević, Alexander M. Petersen, Filippo Radicchi, Roberta Sinatra, Brian Uzzi, Alessandro Vespignani, Ludo Waltman, Dashun Wang, and Albert-László Barabási. 2018. Science of science. *Science*, 359(6379):eaao0185.

Jacob G. Foster, Andrey Rzhetsky, and James A. Evans. 2015. Tradition and innovation in scientists' research strategies. *American Sociological Review*, 80(5):875–908.

Benjamin Freeling, Zoë A. Doubleday, and Sean D. Connell. 2019. How can we boost the impact of publications? Try better writing. *Proceedings of the National Academy of Sciences*, 116(2):341–343.

Benjamin S. Freeling, Zoë A. Doubleday, Matthew J. Dry, Carolyn Semmler, and Sean D. Connell. 2021. Better writing in scientific publications builds reader confidence and understanding. *Frontiers in Psychology*, 12.

Matt Gardner, William Merrill, Jesse Dodge, Matthew Peters, Alexis Ross, Sameer Singh, and Noah A. Smith. 2021. Competency problems: On finding and removing artifacts in language data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1801–1813, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Aaron Gerow, Yuening Hu, Jordan Boyd-Graber, David M. Blei, and James A. Evans. 2018. Measuring discursive influence across scholarship. *Proceedings of the National Academy of Sciences*, 115(13):3308–3313.

Gregorio González-Alcaide, Juan Carlos Valderrama-Zurián, and Rafael Aleixandre-Benavent. 2012. The impact factor in non-English-speaking countries. *Scientometrics*, 92(2):297–311.

Michael Gusenbauer and Neal R. Haddaway. 2020. Which academic search systems are suitable for systematic reviews or meta-analyses? Evaluating retrieval qualities of Google Scholar, PubMed, and 26 other resources. *Research Synthesis Methods*, 11(2):181–217.

Andrew Head, Kyle Lo, Dongyeop Kang, Raymond Fok, Sam Skjonsberg, Daniel S. Weld, and Marti A. Hearst. 2021. Augmenting scientific papers with just-in-time, position-sensitive definitions of terms and symbols. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA. Association for Computing Machinery.

Joseph M Hilbe. 2011. *Negative binomial regression*. Cambridge University Press.

Bas Hofstra, Vivek V. Kulkarni, Sebastian Munoz-Najar Galvez, Bryan He, Dan Jurafsky, and Daniel A. McFarland. 2020. The diversity-innovation paradox in science. *Proceedings of the National Academy of Sciences*, 117(17):9284–9291.

Zhi Hong, Aswathy Ajith, Gregory Pauloski, Eamon Duede, Carl Malamud, Roger Magoulas, Kyle Chard, and Ian Foster. 2022. ScholarBERT: Bigger is not always better.

David Jurgens, Srijan Kumar, Raine Hoover, Dan McFarland, and Dan Jurafsky. 2018. Measuring the evolution of a scientific field through citation frames. *Transactions of the Association for Computational Linguistics*, 6:391–406.

Yea-Seul Kim, Jessica Hullman, Matthew Burgess, and Eytan Adar. 2016. SimpleScience: Lexical simplification of scientific terminology. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1066–1071, Austin, Texas. Association for Computational Linguistics.

Ann Koopman. 2011. Nature launches new open access journal: Scientific reports. *Thomas Jefferson University Library News*.

William Labov. 1973. *Sociolinguistic patterns*. 4. University of Pennsylvania Press.

Vincent Larivière and Yves Gingras. 2010. On the relationship between interdisciplinarity and scientific impact. *Journal of the American Society for Information Science and Technology*, 61(1):126–131.

Omer Levy, Yoav Goldberg, and Ido Dagan. 2015. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225.

Loet Leydesdorff, Caroline S. Wagner, and Lutz Bornmann. 2019. Interdisciplinarity as diversity in citation patterns among journals: Rao-stirling diversity, relative variety, and the gini coefficient. *Journal of Informetrics*, 13(1):255–269.

Liming Liang, Ronald Rousseau, and Zhen Zhong. 2013. Non-English journals and papers in physics and chemistry: Bias in citations? *Scientometrics*, 95(1):333–350.

Fang Liu, Guangyuan Hu, Li Tang, and Weishu Liu. 2018. The penalty of containing more non-English articles. *Scientometrics*, 114(1):359–366.

Yang Liu, Alan Medlar, and Dorota Głowacka. 2022. Lexical ambiguity detection in professional discourse. *Information Processing & Management*, 59(5):103000.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. 2020. S2ORC: The semantic scholar open research corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4969–4983, Online. Association for Computational Linguistics.

Li Lucy and David Bamman. 2021. Characterizing English variation across social media communities with BERT. *Transactions of the Association for Computational Linguistics*, 9:538–556.

Marco Lui and Timothy Baldwin. 2012. langid.py: An off-the-shelf language identification tool. In *Proceedings of the ACL 2012 System Demonstrations*, pages 25–30, Jeju Island, Korea. Association for Computational Linguistics.

Alejandro Martínez and Stefano Mammola. 2021. Specialized terminology reduces the number of citations of scientific papers. *Proceedings of the Royal Society B*, 288(1948):20202581.

Kathy McKeown, Hal Daume III, Snigdha Chaturvedi, John Paparrizos, Kapil Thadani, Pablo Barrio, Or Biran, Suvarna Bothe, Michael Collins, Kenneth R. Fleischmann, Luis Gravano, Rahul Jha, Ben King, Kevin McInerney, Taesun Moon, Arvind Neelakantan, Diarmuid O'Seaghdha, Dragomir Radev, Clay Templeton, and Simone Teufel. 2016. Predicting the impact of scientific concepts using full-text features. *Journal of the Association for Information Science and Technology*, 67(11):2684–2696.

Peter McMahan and James Evans. 2018. Ambiguity and engagement. *American Journal of Sociology*, 124(3):860–912.

Mostafa Mesgari, Chitu Okoli, Mohamad Mehdi, Finn Årup Nielsen, and Arto Lanamäki. 2015. "the sum of all human knowledge": A systematic review of scholarly research on the content of w ikipedia. *Journal of the Association for Information Science and Technology*, 66(2):219–245.

Sonia K Murthy, Kyle Lo, Daniel King, Chandra Bhagavatula, Bailey Kuehl, Sophie Johnson, Jonathan Borchardt, Daniel S Weld, Tom Hope, and Doug Downey. 2022. Accord: A multi-document approach to generating diverse descriptions of scientific concepts. In *Proceedings of the EMNLP 2022 System Demonstrations*.

Innocent Ndubuisi-Obi, Sayan Ghosh, and David Jurgens. 2019. Wetin dey with these comments? modeling sociolinguistic factors affecting code-switching behavior in nigerian online discussions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6204–6214, Florence, Italy. Association for Computational Linguistics.

Mark EJ Newman. 2016. Equivalence between modularity optimization and maximum likelihood methods for community detection. *Physical Review E*, 94(5):052315.

Sing Bik Cindy Ngai, Rita Gill Singh, and Alex Chun Koon. 2018. A discourse analysis of the macro-structure, metadiscoursal and microdiscoursal features in the abstracts of research articles across multiple science disciplines. *PLOS ONE*, 13(10):1–21.

Dong Nguyen, A. Seza Doğruöz, Carolyn P. Rosé, and Franciska de Jong. 2016. Survey: Computational sociolinguistics: A Survey. *Computational Linguistics*, 42(3):537–593.

Keisuke Okamura. 2019. Interdisciplinarity revisited: evidence for research impact and dynamism. *Palgrave Communications*, 5(1):1–9.

José Luis Ortega and Isidro F. Aguillo. 2014. Microsoft academic search and google scholar citations: Comparative analysis of author profiles. *Journal of the Association for Information Science and Technology*, 65(6):1149–1156.

Umashanthi Pavalanathan and Jacob Eisenstein. 2015. Audience-Modulated Variation in Online Social Media. *American Speech*, 90(2):187–213.

Hao Peng, Qing Ke, Ceren Budak, Daniel M Romero, and Yong-Yeol Ahn. 2021. Neural embeddings of scholarly periodicals reveal complex disciplinary organizations. *Science Advances*, 7(17):eabb9004.

Pontus Plavén-Sigray, Granville James Matheson, Björn Christian Schiffler, and William Hedley Thompson. 2017. Research: The readability of scientific texts is decreasing over time. *eLife*, 6:e27725.

Vinodkumar Prabhakaran, William L. Hamilton, Dan McFarland, and Dan Jurafsky. 2016. Predicting the rise and fall of scientific topics from trends in their rhetorical framing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1170–1180, Berlin, Germany. Association for Computational Linguistics.

Tzipora Rakedzon, Elad Segev, Noam Chapnik, Roy Yosef, and Ayelet Baram-Tsabari. 2017. Automatic jargon identifier for scientists engaging with the public and science communication educators. *PLOS ONE*, 12(8):1–13.

Abhinav Ramesh Kashyap, Devamanyu Hazarika, Min-Yen Kan, and Roger Zimmermann. 2021. Domain divergences: A survey and empirical analysis. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1830–1849, Online. Association for Computational Linguistics.

Joseph Reagle and Jackie Koerner. 2020. *Wikipedia@ 20: Stories of an incomplete revolution*. The MIT Press.

Martin Rosvall and Carl T Bergstrom. 2008. Maps of random walks on complex networks reveal community structure. *Proceedings of the national academy of sciences*, 105(4):1118–1123.

Janet N. Ryan. 1985. The language gap: Common words with technical meanings. *Journal of Chemical Education*, 62(12):1098.

B. C. Satishkumar, P. John Thomas, A. Govindaraj, and C. N. R. Rao. 2000. Y-junction carbon nanotubes. *Applied Physics Letters*, 77(16):2530.

Yanchuan Sim, Noah A. Smith, and David A. Smith. 2012. Discovering factions in the computational linguistics community. In *Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries*, pages 22–32, Jeju Island, Korea. Association for Computational Linguistics.

Arnab Sinha, Zhihong Shen, Yang Song, Hao Ma, Darrin Eide, Bo-June (Paul) Hsu, and Kuansan Wang. 2015. An overview of microsoft academic service (mas) and applications. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15 Companion, page 243–246, New York, NY, USA. Association for Computing Machinery.

Sandeep Soni, David Bamman, and Jacob Eisenstein. 2022. Predicting long-term citations from short-term linguistic influence.

Sandeep Soni, Kristina Lerman, and Jacob Eisenstein. 2021. Follow the leader: Documents on the leading edge of semantic change get more citations. *Journal of the Association for Information Science and Technology*, 72(4):478–492.

Andrew Stirling. 1998. On the economics and analysis of diversity. *Science Policy Research Unit (SPRU), Electronic Working Papers Series, Paper*, 28:1–156.

Raghuram Vadapalli, Bakhtiyar Syed, Nishant Prabhu, Balaji Vasan Srinivasan, and Vasudeva Varma. 2018. When science journalism meets artificial intelligence : An interactive demonstration. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 163–168, Brussels, Belgium. Association for Computational Linguistics.

Richard Van Noorden. 2015. Interdisciplinary research by the numbers. *Nature*, 525(7569):306–307.

Harold Varmus, Patrick Brown, and Michael Eisen. 2000. Open letter. *PLOS One*.

Daril A Vilhena, Jacob G Foster, Martin Rosvall, Jevin D West, James Evans, and Carl T Bergstrom. 2014. Finding cultural holes: How structure and culture diverge in networks of scholarly communication. *Sociological Science*, 1:221.

Dashun Wang and Albert-László Barabási. 2021. *The h-Index*, page 17–27. Cambridge University Press.

Kuansan Wang, Zhihong Shen, Chiyuan Huang, Chieh-Han Wu, Darrin Eide, Yuxiao Dong, Junjie Qian, Anshul Kanakia, Alvin Chen, and Richard Rogahn. 2019. A review of Microsoft Academic Services for science of science studies. *Frontiers in Big Data*, 2.

Jevin West and Jason Portenoy. 2016. Delineating fields using mathematical jargon. In *Proceedings of the*

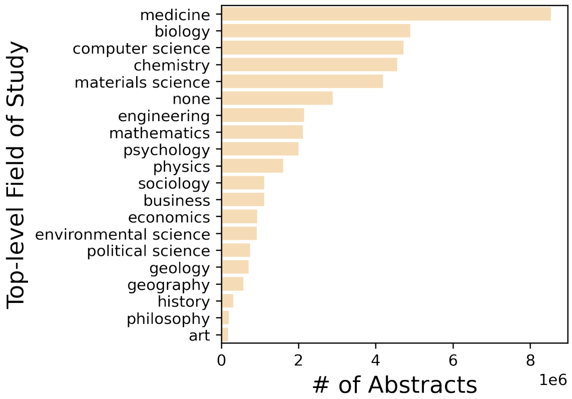


Figure 7: The number of abstracts in each top-level MAG field of study (FOS) in S2ORC, sorted from biggest to smallest, before subsampling. Medicine is the most frequent field, and is almost twice as large as the next largest field, which is biology.

Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL), pages 63–71.

Dani Yogatama, Michael Heilman, Brendan O'Connor, Chris Dyer, Bryan R. Routledge, and Noah A. Smith. 2011. Predicting a scientific community's response to an article. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 594–604, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Justine Zhang, William Hamilton, Cristian Danescu-Niculescu-Mizil, Dan Jurafsky, and Jure Leskovec. 2017. Community identity and user engagement in a multi-community landscape. *Proceedings of the International AAAI Conference on Web and Social Media*, 11(1):377–386.

A Fields of study

Figure 7 shows the number of valid abstracts in each top-level MAG field of study *before* subsampling a similar number of abstracts from each subfield. This figure can be compared to Figure 2 to show how the distribution of fields changed after sampling. The following lists the subfields, or level 1 MAG FOS, used in our study. The same subfield may fall under multiple fields.

- **Art** (6 children): art history, classics, humanities, visual arts, literature, aesthetics.

- **Biology** (31 children): computational biology, biochemistry, bioinformatics, cancer research, evolutionary biology, anatomy, molecular biology, pharmacology, immunology, virology, ecology, agronomy, botany, toxicology, food science, microbiology, biological system, agroforestry, biophysics, animal science, paleontology, cell biology, physiology, endocrinology, horticulture, genetics, biotechnology, neuroscience, fishery, zoology, biology (other).

- **Business** (10 children): international trade, accounting, risk analysis (engineering), process management, actuarial science, marketing, industrial organization, finance, advertising, business (other).

- **Chemistry** (20 children): polymer chemistry, molecular physics, biochemistry, organic chemistry, physical chemistry, chemical physics, nuclear chemistry, medicinal chemistry, photochemistry, combinatorial chemistry, computational chemistry, analytical chemistry, food science, chromatography, mineralogy, inorganic chemistry, crystallography, stereochemistry, environmental chemistry, chemistry (other).

- **Computer Science** (32 children): natural language processing, software engineering, theoretical computer science, embedded system, computer security, programming language, data science, computer vision, computer network, human–computer interaction, world wide web, information retrieval, parallel computing, operating system, computer hardware, multimedia, computer graphics (images), library science, real-time computing, artificial intelligence, database, distributed computing, simulation, telecommunications, internet privacy, pattern recognition, machine learning, knowledge management, data mining, speech recognition, algorithm, computer science (other).

- **Economics** (28 children): international trade, labour economics, political economy, natural resource economics, industrial organization, monetary economics, economic system, economy, operations management, demographic economics, management, finance, management science, environmental resource management, accounting, agricultural economics, economic growth, actuarial science, financial economics, market economy, socioeconomics, environmental economics, econometrics, law and economics, development economics, public economics, microeconomics, economics (other).

- **Engineering** (35 children): engineering ethics, software engineering, control engineering, embedded system, nuclear engineering, reliability engineering, operations research, transport engineering, engineering drawing, biomedical engineering, engineering management, electronic engineering, automotive engineering, forensic engineering, operations management, mechanical engineering, petroleum engineering, process engineering, systems engineering, management science, civil engineering, control theory, simulation, telecommunications, geotechnical engineering, pulp and paper industry, process management, environmental engineering, marine engineering, chemical engineering, manufacturing engineering, waste management, structural engineering, electrical engineering, engineering (other).

- **Environmental Science** (7 children): environmental resource management, environmental planning, environmental engineering, agroforestry, soil science, environmental protection, environmental science (other).

- **Geography** (7 children): environmental planning, meteorology, archaeology, physical geography, remote sensing, environmental protection, geography (other).

- **Geology** (14 children): atmospheric sciences, geochemistry, geomorphology, soil science, hydrology, oceanography, climatology, mineralogy, geotechnical engineering, seismology, petroleum engineering, remote sensing, paleontology, geology (other).

- **History** (5 children): art history, classics, ancient history, archaeology, history (other).

- **Materials Science** (6 children): polymer chemistry, optoelectronics, composite material, nanotechnology, metallurgy, materials science (other).

- **Mathematics** (17 children): geometry, topology, combinatorics, operations research, mathematical optimization, pure mathematics, control theory, discrete mathematics, statistics, algebra, mathematics education, mathematical physics, applied mathematics, econometrics, mathematical analysis, algorithm, mathematics (other).

- **Medicine** (45 children): audiology, gerontology, pediatrics, obstetrics, medical physics, urology, radiology, gynecology, dentistry, cancer research, cardiology, veterinary medicine, biomedical engineering, medical education, general surgery, andrology, oncology, dermatology, traditional medicine, orthodontics, anatomy, pharmacology, medical emergency, anesthesia, gastroenterology, immunology, virology, risk analysis (engineering), emergency medicine, surgery, psychiatry, physiology, nursing, endocrinology, clinical psychology, intensive care medicine, physical therapy, nuclear medicine, family medicine, ophthalmology, environmental health, internal medicine, physical medicine and rehabilitation, pathology, medicine (other).

- **Philosophy** (6 children): environmental ethics, humanities, epistemology, aesthetics, linguistics, philosophy (other).

- **Physics** (24 children): mechanics, atmospheric sciences, molecular physics, astrophysics, acoustics, medical physics, classical mechanics, chemical physics, nuclear physics, optoelectronics, quantum mechanics, theoretical physics, optics, computational physics, particle physics, atomic physics, statistical physics, meteorology, nuclear magnetic resonance, thermodynamics, mathematical physics, astronomy, condensed matter physics, physics (other).

- **Political Science** (4 children): public relations, public administration, law, political science (other).

- **Psychology** (15 children): mathematics education, cognitive psychology, criminology, clinical psychology, applied psychology, social psychology, communication, pedagogy, psychoanalysis, neuroscience, developmental psychology, psychiatry, psychotherapist, cognitive science, psychology (other).

- **Sociology** (11 children): social science, criminology, demography, law and economics, communication, pedagogy, political economy, gender studies, socioeconomics, media studies, sociology (other).

B Dataset filtering

We perform the following preprocessing steps of S2ORC to create CONTEMPORARY S2ORC:

- **Venue.** We consolidate the *venue* and *journal* keys of each article's metadata. We use whichever label is non-empty, and only a small fraction (0.08%) of articles with valid abstracts have *venue* and *journal* that differ, in which case we use use the article's *journal*. We handle venue names case insensitively, and also remove tokens in their names that contain numbers to consolidate years and editions.

- **Time**. Our study focuses on contemporary science, which are abstracts published during 2000-2019. S2ORC contains some abstracts from 2020 and onwards, but dates past 2020 are likely metadata processing errors. We remove 47.6 million articles outside of this time range.

- **Valid metadata**. We remove 42.5 million articles with missing abstracts, titles, or journal and venue labels.

- **Language.** We remove 77,133 articles from 925 non-English journals or venues, which are those that have less than 80% of their articles in English, using Lui and Baldwin (2012)'s language classifier.

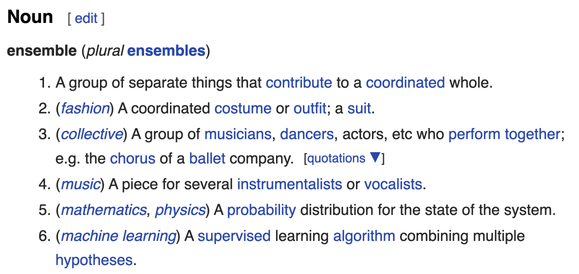
- **Field of study.** Medicine fields dominate S2ORC abstracts. We balance the dataset by taking a sample of 50k articles per subfield. For subfields that are too small to sample or articles that have field-level but no subfield-level labels, we categorize these in an OTHER subfield under their parent field. Since articles can be labeled with multiple FOS, our sample is not perfectly stratified, but prevents large subfields from dominating calculations of the general prevalence of words in English. In total we identify specialized language across 293 subfields that fall under 19 fields (listed in Appendix A).

C Validation details

Here, we include two additional figures to supplement §4.

Figure 8 shows a screenshot of a Wiktionary entry for the word *ensemble*, which is overloaded with several labeled definitions.[10] Some labels, such as *collective*, show grammatical information, while others indicate restricted usage to different fields, dialects, or contexts.[11] We match these labels to MAG fields and subfields when evaluating recall of words marked as discipline-specific by Wiktionary.

In the main text, we show that sense NPMI is able to recall more Wiktionary words at the same threshold than type NPMI. In addition, sense NPMI scores are higher in Wiktionary-labeled fields than random ones (Figure 9).



Noun [edit]

ensemble (*plural* **ensembles**)

1. A group of separate things that contribute to a coordinated whole.
2. (*fashion*) A coordinated costume or outfit; a suit.
3. (*collective*) A group of musicians, dancers, actors, etc who perform together; e.g. the chorus of a ballet company. [quotations ▼]
4. (*music*) A piece for several instrumentalists or vocalists.
5. (*mathematics*, *physics*) A probability distribution for the state of the system.
6. (*machine learning*) A supervised learning algorithm combining multiple hypotheses.

Figure 8: An example of a Wiktionary entry for the word *ensemble*. This word has definitions labeled as pertaining to *fashion* (a coordinated outfit), *collective* (a group of performers), *music* (a musical piece), *mathematics & physics* (a probability distribution), and *machine learning* (a supervised learning algorithm).

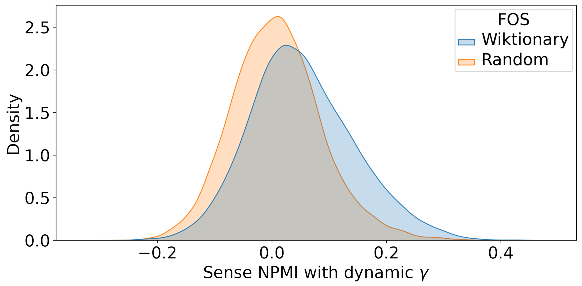


Figure 9: The distribution of sense NPMI scores for words in Wiktionary-labeled fields versus random ones. Words labeled as belonging to a subfield by Wiktionary have higher $\mathcal{S}_f(t)$ in that subfield than in a random one (paired t-test, $p < 0.001$).

D Additional experimental details

D.1 Cutoff decision

We generated Figure 5 with additional values of the NPMI cutoff c, such as $c = 0.2$, and achieve similar conclusions (Figure 10). That is, these results are similar when it comes to which fields tend to adjust their language between general-purpose and discipline-focused venues. In the main text, we

[10] https://en.wiktionary.org/wiki/ensemble

[11] https://en.wiktionary.org/wiki/Template:label

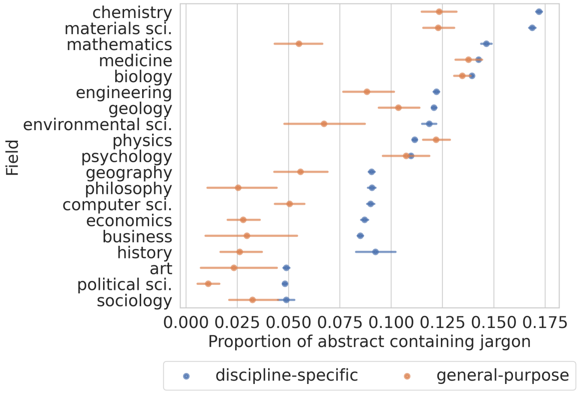


Figure 10: Abstracts' average fraction of discipline-specific words and senses in fields that appear in both general-purpose and discipline-focused venues (95% CI). This plot uses an NPMI cutoff of 0.2.

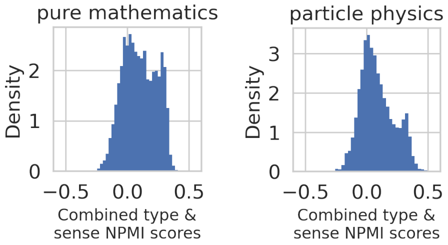


Figure 11: Sometimes, NPMI score distributions for subfields are bimodal with a second peak among positive values, especially when a subfield contains large amounts of jargon. The left shows the distribution for pure mathematics, while the right shows particle physics.

usually use $c = 0.1$, as positive NPMI values indicate association, but NPMI values too close to 0 would instead lean towards independence. Though NPMI ranges from -1 to 1, the outputted scores for various subfields tended to range from -0.5 to 0.5, and some include bimodal behavior where the latter peak of the distribution usually occurs after $c = 0.2$ (Figure 11). We assume that this latter peak is indicative of jargon. Thus, we experimented with cutoffs that would separate the initial peak around 0 and a secondary peak in the positive NPMI value range, if any.

D.2 Scholarly success

D.2.1 Subfield similarity

To calculate subfield similarity, we first create a $(N+1) \times (N+1)$ citation matrix, where N is the total number of subfields, and the additional row and column represents articles in unknown subfields. Rows in this matrix represent subfields that are cited, and columns are citing subfields. This matrix is generated using all articles published in S2ORC within the years 2000 and 2019 that have inbound citations. For subfield similarity calculations, we use the rows to represent each subfield. For example, the nearest neighbors via cosine similarity of the row representing *chemical engineering* include *polymer chemistry*, *polymer science*, and *inorganic chemistry*.

D.2.2 Regressions

We ran a few statistical tests to determine what regressions to use.

Citation counts. We run both Poisson regressions and negative binomial regressions on citation count data, as these generalized linear models are typically used to model count data. Negative binomial regression is used for data that shows overdispersion, when the variance of the dependent variable exceeds the mean. We calculate the overdispersion ratio ϕ of Poisson regressions for each field:

$$\phi = \frac{\text{Pearson's } \chi^2}{\text{residual degrees of freedom}}.$$

Since it exceeds 1 for each field's regression, there is overdispersion in our data, and thus we use negative binomial regressions for citation counts. Negative binomial regressions require choosing a constant α which is used to express the variance in terms of the mean. We determine α by inputting the fitted rate vector from the Poisson regression into an auxiliary OLS regression without a constant (Cameron and Trivedi, 2013). The α we obtain from this for each regression is significant for all fields except for Art and Philosophy ($p < 0.01$, right-tailed t-test).

Interdisciplinary impact. We run ordinary least squares (OLS) regressions for this dependent variable. OLS involves several assumptions: randomly sampled data, linearity, exogeneity, noncollinearity, and homoskedasticity. We check for linearity and exogeneity by comparing residuals and fitted values, non-collinearity by checking that the variance inflation factors of covariates do not exceed 5, and homoskedasticity by running a Breusch-Pagan test (Breusch and Pagan, 1979). We find that we satisfy all assumptions except homoskedasticity. Due to to this, we also run a weighted least squares regression to check the robustness of our OLS results, and achieve similar coefficients.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Section 9

☑ A2. Did you discuss any potential risks of your work?
Section 10

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Section 1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

1

☑ B1. Did you cite the creators of artifacts you used?
2

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
10

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
10

☑ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
10

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
2

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
2 (dataset size), 4 (Wiktionary # of examples), 6 (# of observations and venues in experiments)

C ☑ Did you run computational experiments?

4, 5, 6

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
9

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
4

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
4, 5, 6

☒ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
We use a word sense induction model that isn't an existing package, but was open-source, and we detailed any changes and parameter settings we used for that model.

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

□ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

□ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

□ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

□ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

□ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.