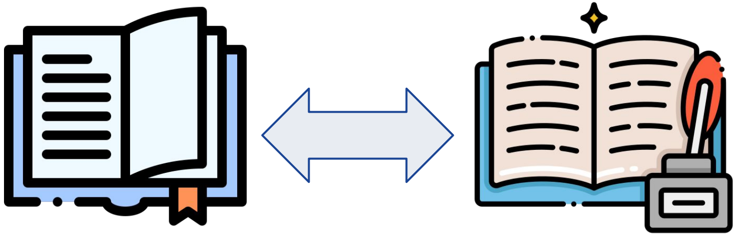
Words as Gatekeepers: Measuring Discipline-specific Terms and Meanings in Scholarly Publications

Li Lucy, Jesse Dodge, David Bamman, **Katherine A. Keith**

New Directions in Analyzing Text as Data (TADA)

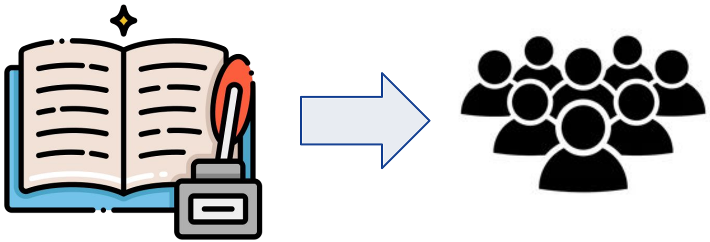
Previous work on scholarly language

Between fields



McKeown et al., 2016; Prabhakaran et al., 2016; Sim et al., 2012; Rakedzon et al., 2017, Vilhena et al. 2014

Between scientific communities and the public

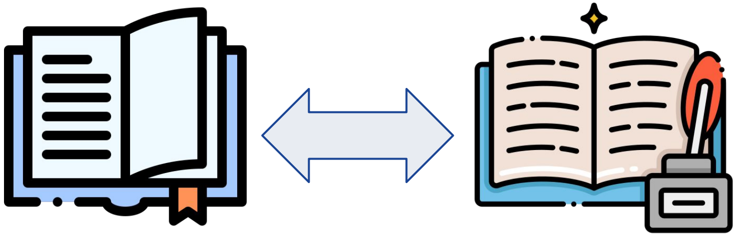


Liu et al., 2022; August et al., 2020a; Cervetti et al., 2015; Freeling et al. 2021

Scholars to dogs … not so much

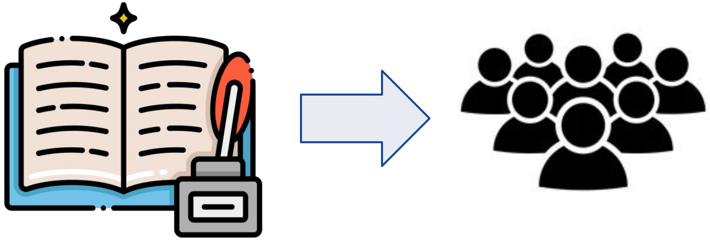
Previous work on scholarly language

Between fields



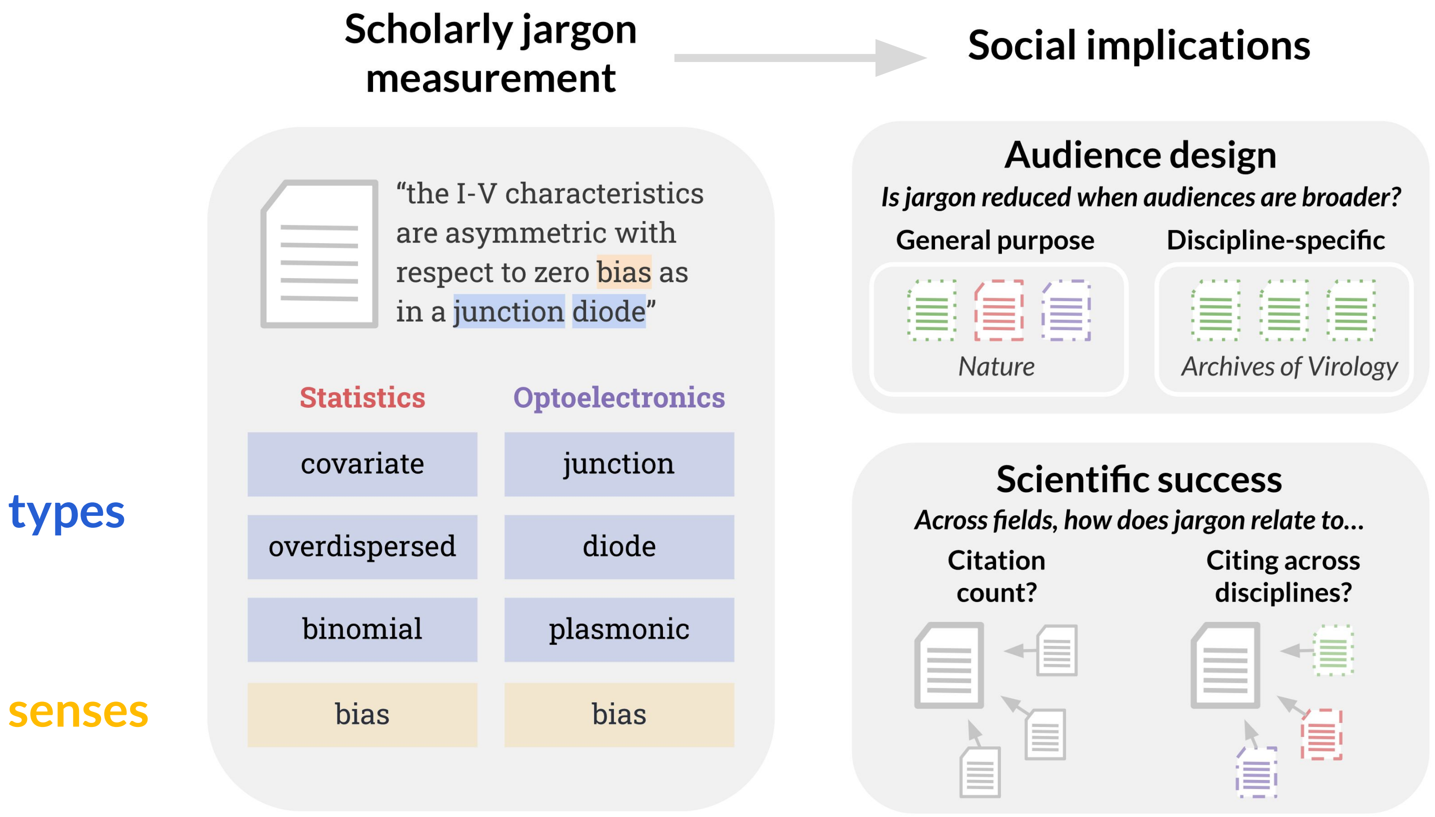
McKeown et al., 2016; Prabhakaran et al., 2016; Sim et al., 2012; Rakedzon et al., 2017, Vilhena et al. 2014

Between scientific communities and the public



Liu et al., 2022; August et al., 2020a; Cervetti et al., 2015; Freeling et al. 2021

Most use measures of word **types** *not* word **senses.**



Scholarly jargon measurement

“the I-V characteristics are asymmetric with respect to zero bias as in a junction diode”

	Statistics	Optoelectronics
types	covariate	junction
	overdispersed	diode
	binomial	plasmonic
senses	bias	bias

Social implications

Audience design

Is jargon reduced when audiences are broader?

General purpose

Nature

Discipline-specific

Archives of Virology

Scientific success

Across fields, how does jargon relate to...

Citation count?

Citing across disciplines?

Ai2

Data

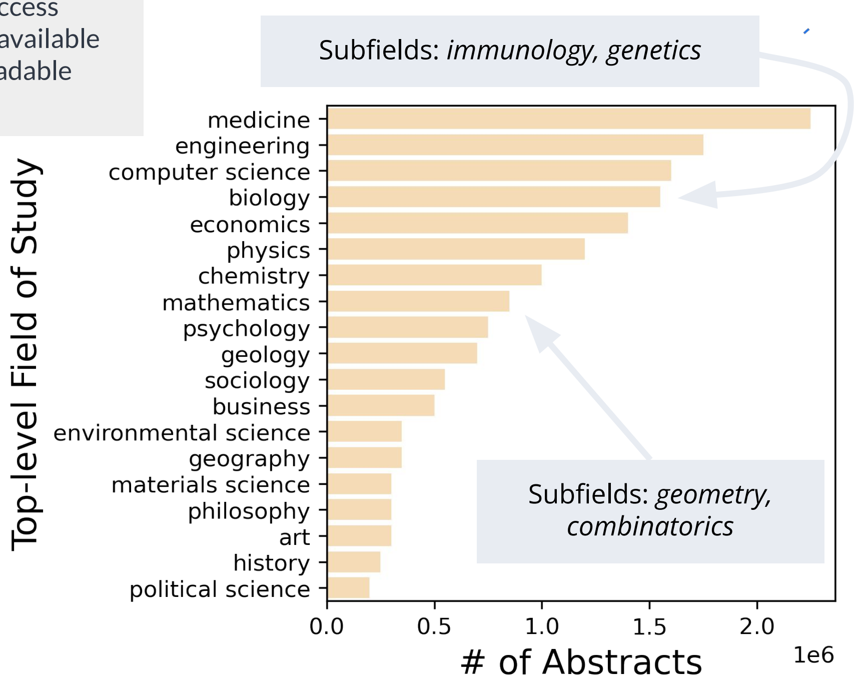
Full text for 8.1M open access papers. Largest publicly-available collection of machine-readable academic text to date

Contemporary S2ORC

- Semantic Scholar Open Research Corpus (S2ORC), English, 2000-2019
- Linked abstracts to 19 fields (with 293 subfields) in the Microsoft Academic Graph
- Sampled so each subfield has same number of abstracts

"Background" corpus

- S2ORC + English Wikipedia sample



Subfields: *immunology, genetics*

Subfields: *geometry, combinatorics*

Discipline-specific word type metric

Normalized pointwise mutual information (NPMI)

“Overall” probability in background corpus

$$\underbrace{\log \frac{P(t \mid s)}{P(t)}}_{\text{PMI of word type } t \text{ in subfield } s} \Big/ \underbrace{-\log P(t, s)}_{\text{Normalizes values to be between -1 and +1}}$$

Also used in Lucy & Bamman (TACL 2021) for community-specific English on the web

Fine print: lemmas, filter to at least 20 instances

Word sense induction

Core idea: Word tokens that share senses also share in-context substitutes

Eyal et al. (ACL 2022)

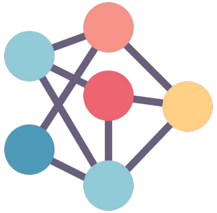
1. Raw text

We used **python**, HTML, CSS, Javascript, node, flask.

→ *slack, oracle, apple, bot, framework*

2. Use ScholarBERT to predict top 5 substitutes of each masked target token

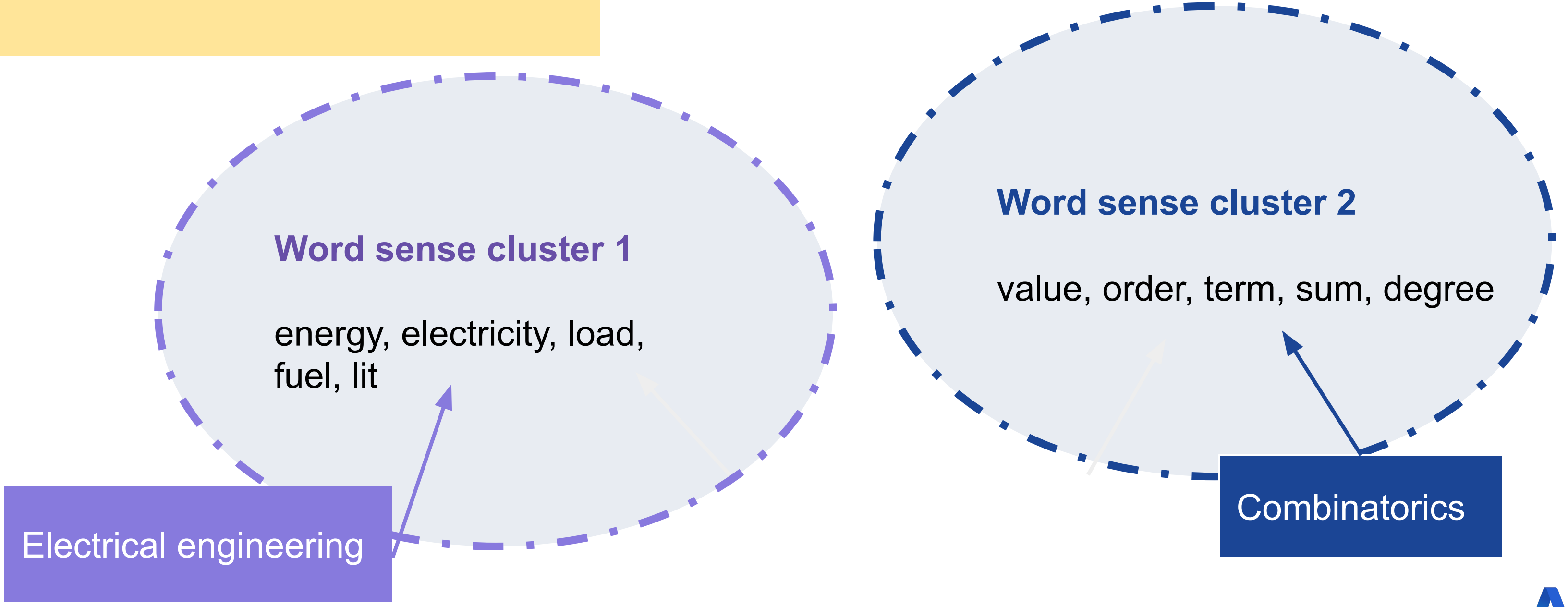
Hong et al. (2022)



3. Make co-occurrence network of a word's substitutes and run community detection algorithm

Example (toy) output

ScholarBERT predictions for masked target word **power**



Word sense cluster 1

energy, electricity, load, fuel, lit

Electrical engineering

Word sense cluster 2

value, order, term, sum, degree

Combinatorics

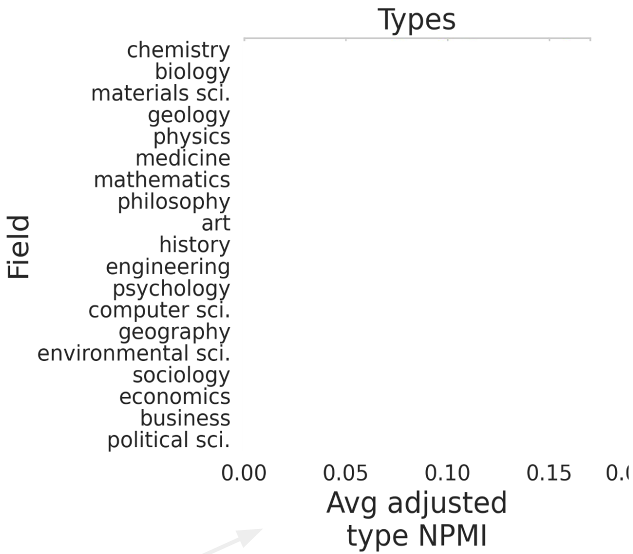
Discipline-specific word sense metric

Normalized pointwise mutual information (NPMI)

Now!
t = word sense cluster id

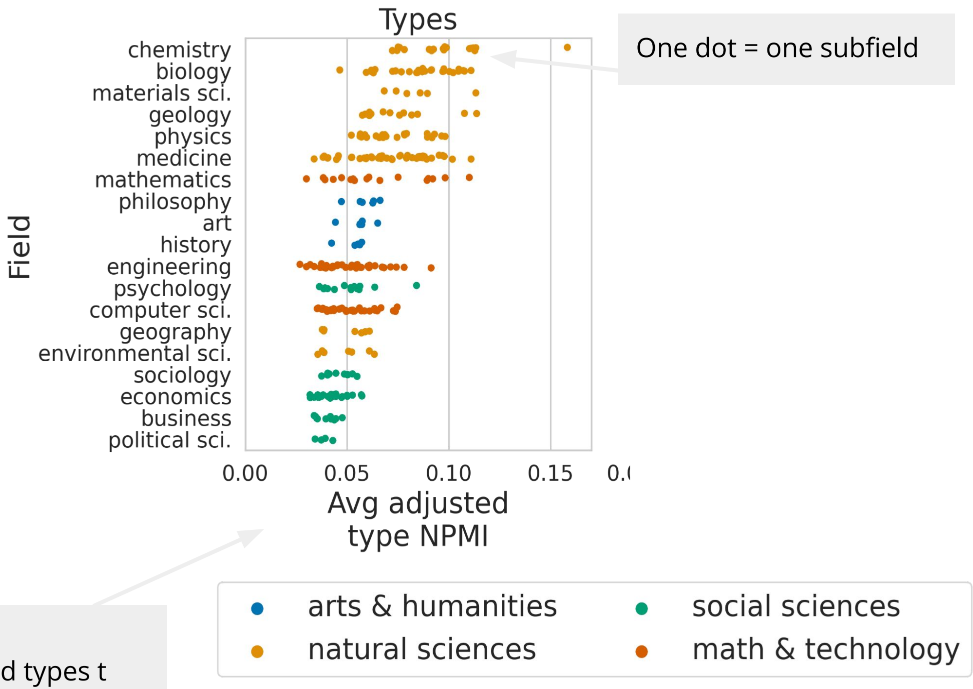
$$\underbrace{\log \frac{P(t \mid s)}{P(t)}}_{\text{PMI of word } \textbf{sense } t \text{ in subfield } s} \Big/ \underbrace{-\log P(t, s)}_{\text{Normalizes values to be between -1 and +1}}$$

Word **types** are distinctive across fields



max(NPMI(s,t), 0)
averaged across all word types t

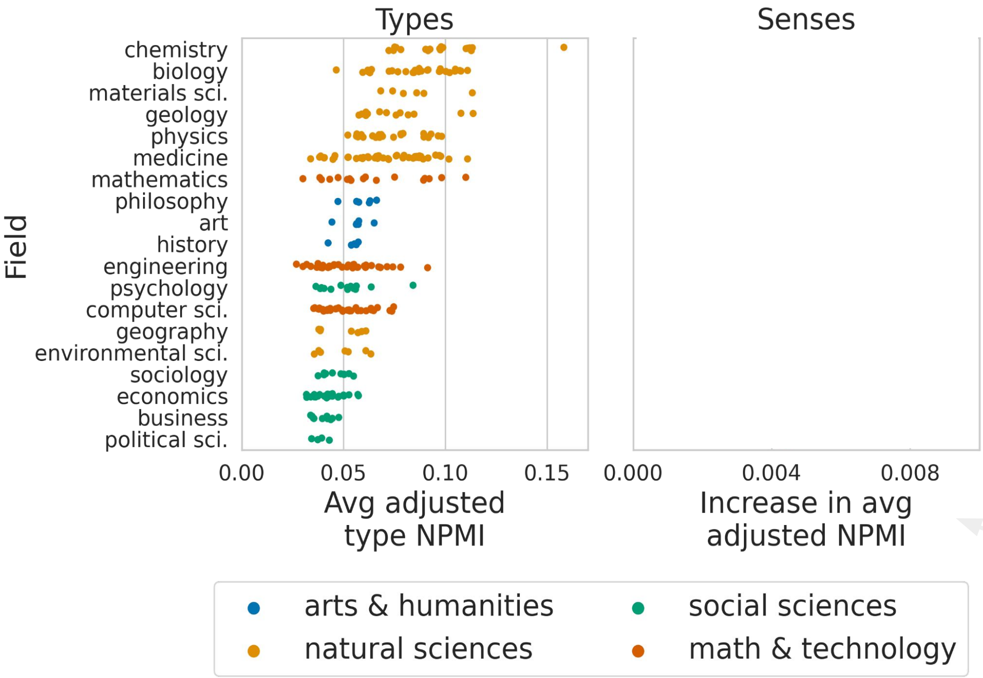
Word types are distinctive across fields



One dot = one subfield

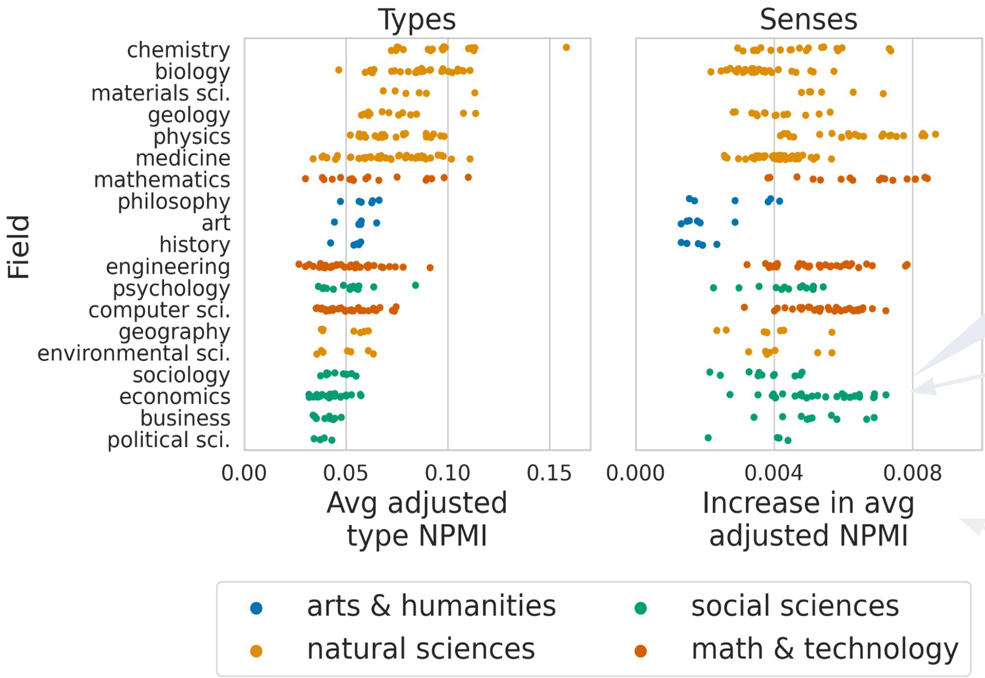
max(NPMI(s,t), 0)
averaged across all word types t

Word senses give a different picture



Increase after
max (sense NPMI, type NPMI)

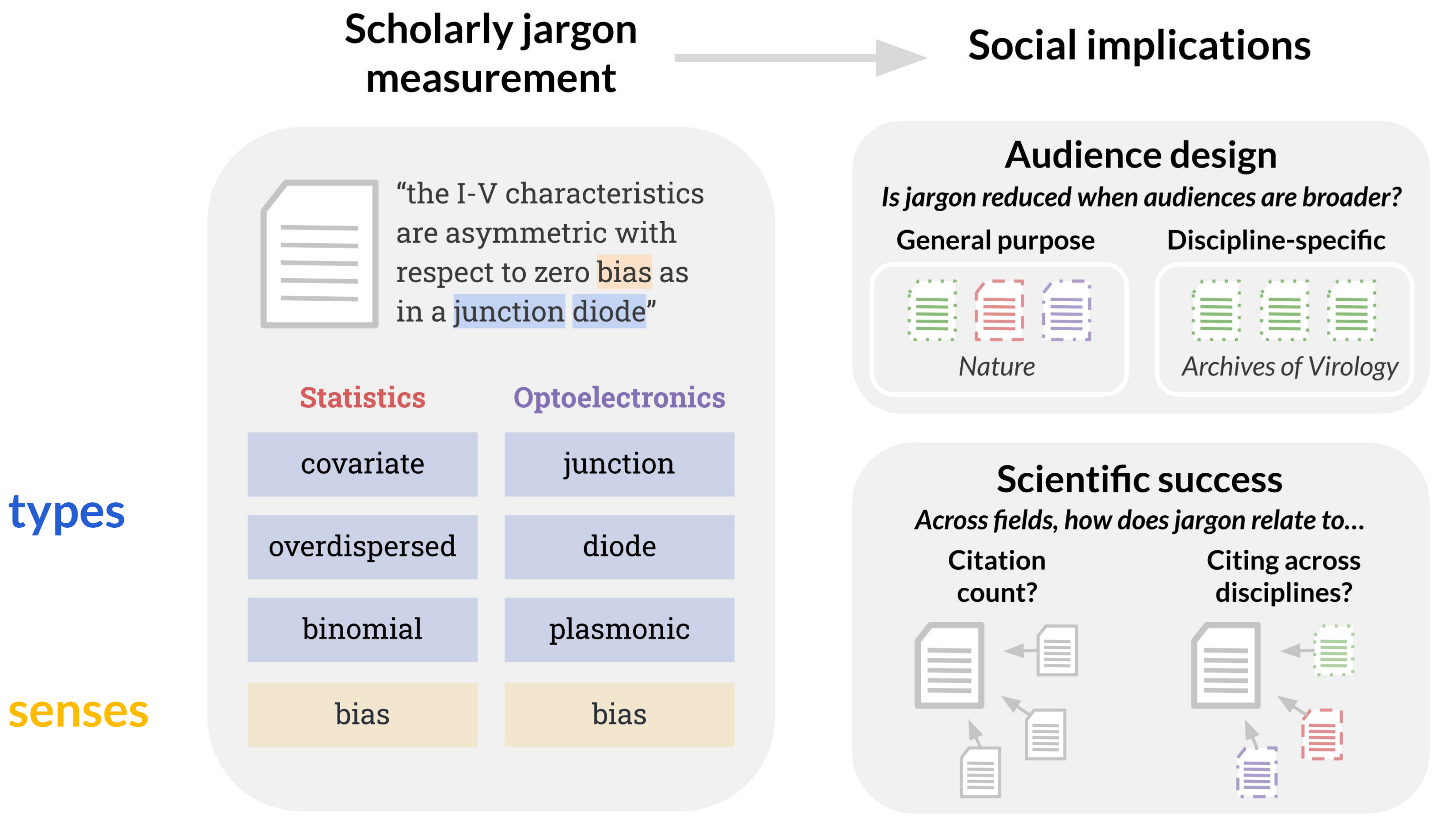
Word senses give a different picture



Takeaway: Ignoring *sense* jargon might collapse complexity of scholarly language in the *social sciences*

Monetary economics
movement, liquid, interest, turbulence, provider

Increase after
max (sense NPMI, type NPMI)



Scholarly jargon measurement → Social implications

“the I-V characteristics are asymmetric with respect to zero bias as in a junction diode”

	Statistics	Optoelectronics
types	covariate	junction
	overdispersed	diode
	binomial	plasmonic
senses	bias	bias

Audience design

Is jargon reduced when audiences are broader?

General purpose — *Nature*

Discipline-specific — *Archives of Virology*

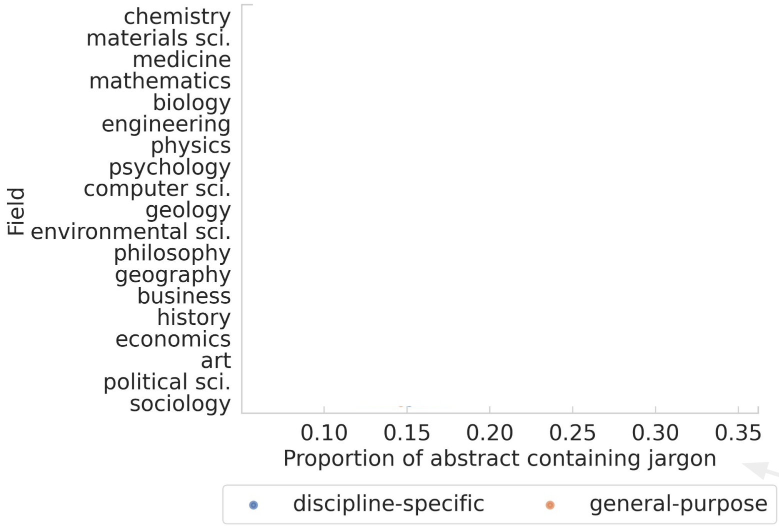
Scientific success

Across fields, how does jargon relate to…

Citation count?

Citing across disciplines?

Audience Design

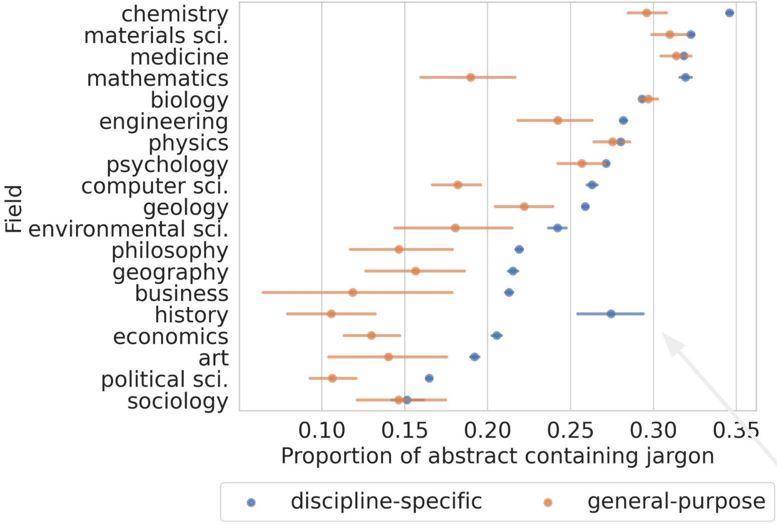


max(sense NPMI, type NPMI) > 0.1

Nature, Nature Communications, PLOS One, Science, Science Advances, and Scientific Reports

Audience Design

Most fields reduce jargon when writing for general-purpose venues

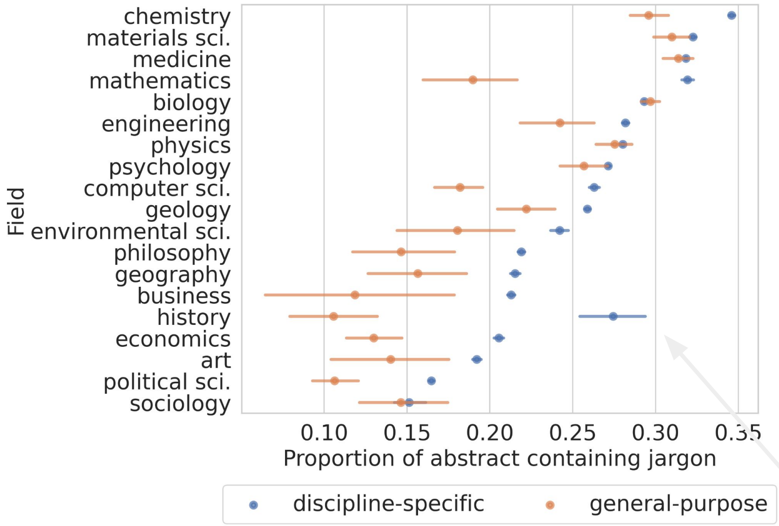


Nature, Nature Communications, PLOS One, Science, Science Advances, and Scientific Reports

95% confidence intervals from bootstrapping

Audience Design

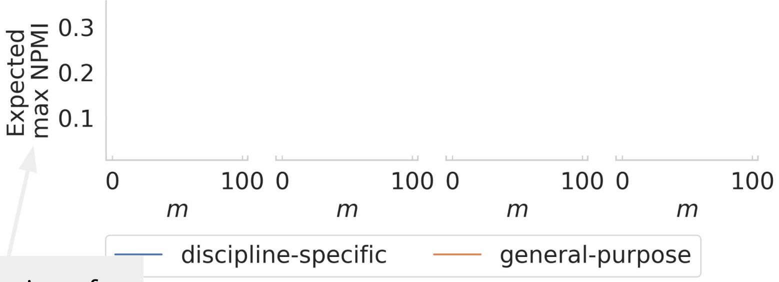
Most fields reduce jargon when writing for general-purpose venues



Nature, Nature Communications, PLOS One, Science, Science Advances, and Scientific Reports

95% confidence intervals from bootstrapping

… but some more than others

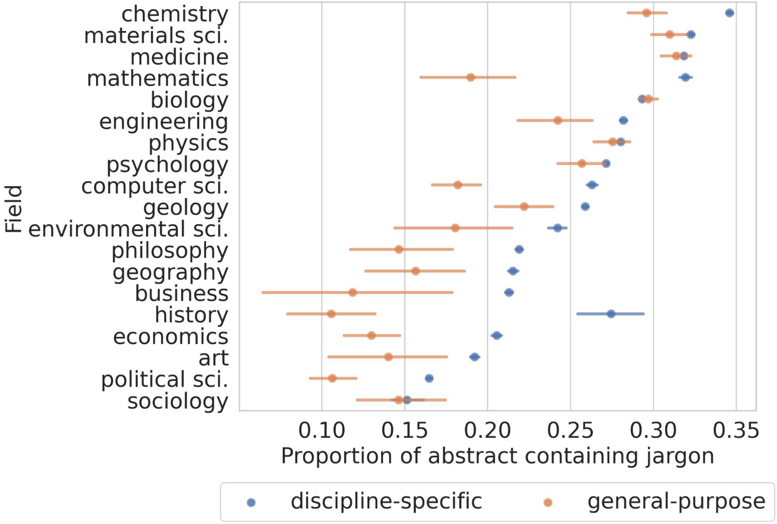


Proportion of max jargon

m = the position of a word in an abstract

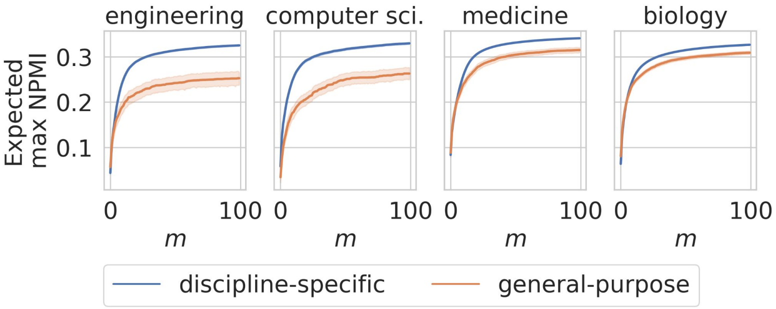
Audience Design

Most fields reduce jargon when writing for general-purpose venues



Nature, Nature Communications, PLOS One, Science, Science Advances, and Scientific Reports

... but some more than others



m = the position of a word in an abstract

Scientific success

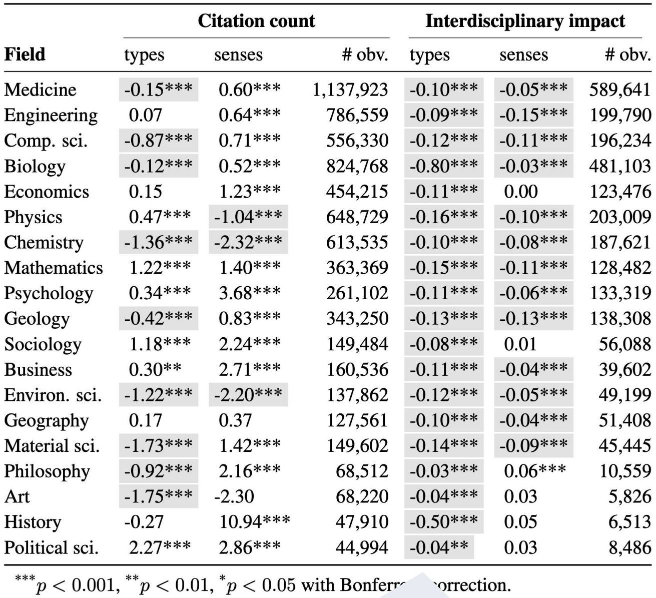
Proportion of Jargon in Abstract ~

1. **Citation count**
   Negative binomial regression

2. **Interdisciplinary impact** Leydesdorff et al. (2019)
   Ordinary least squares regression

$$= \underset{\substack{\uparrow \\ \text{variety}}}{\frac{n}{N}} \underset{\substack{\uparrow \\ \text{balance}}}{(1 - \text{Gini})} \sum_{i,j \in \mathcal{C}, i \neq j} \underset{\substack{\uparrow \\ \text{inter-similarity}}}{\frac{d_{ij}}{n(n-1)}}$$

Other variables in regression: 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		
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 Bonferr orrection.

Takeaway: Increase in jargon almost always *negatively associated* with interdisciplinary impact

More details and analysis in our paper

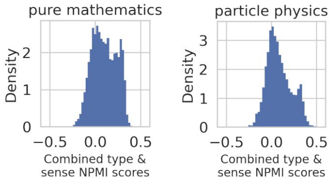


Figure 11: Sometimes, NPMI score distributions for subfields are bimodal with a second peak among positive values, especially when a subfield contains large amounts of jargon. The left shows the distribution for pure mathematics, while the right shows particle physics.

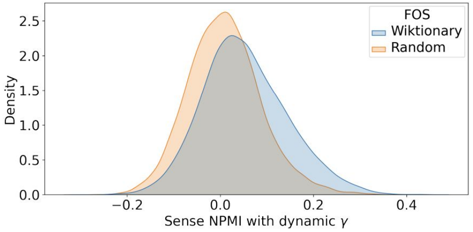


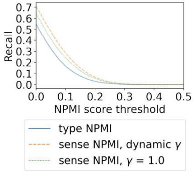
Figure 9: The distribution of sense NPMI scores for words in Wiktionary-labeled fields versus random ones. Words labeled as belonging to a subfield by Wiktionary have higher $\mathcal{S}_f(t)$ in that subfield than in a random one (paired t-test, $p < 0.001$).

	sense t_1			sense t_2		
word t	FOS a	$\mathcal{S}_a(t_1)$	top substitutes	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.

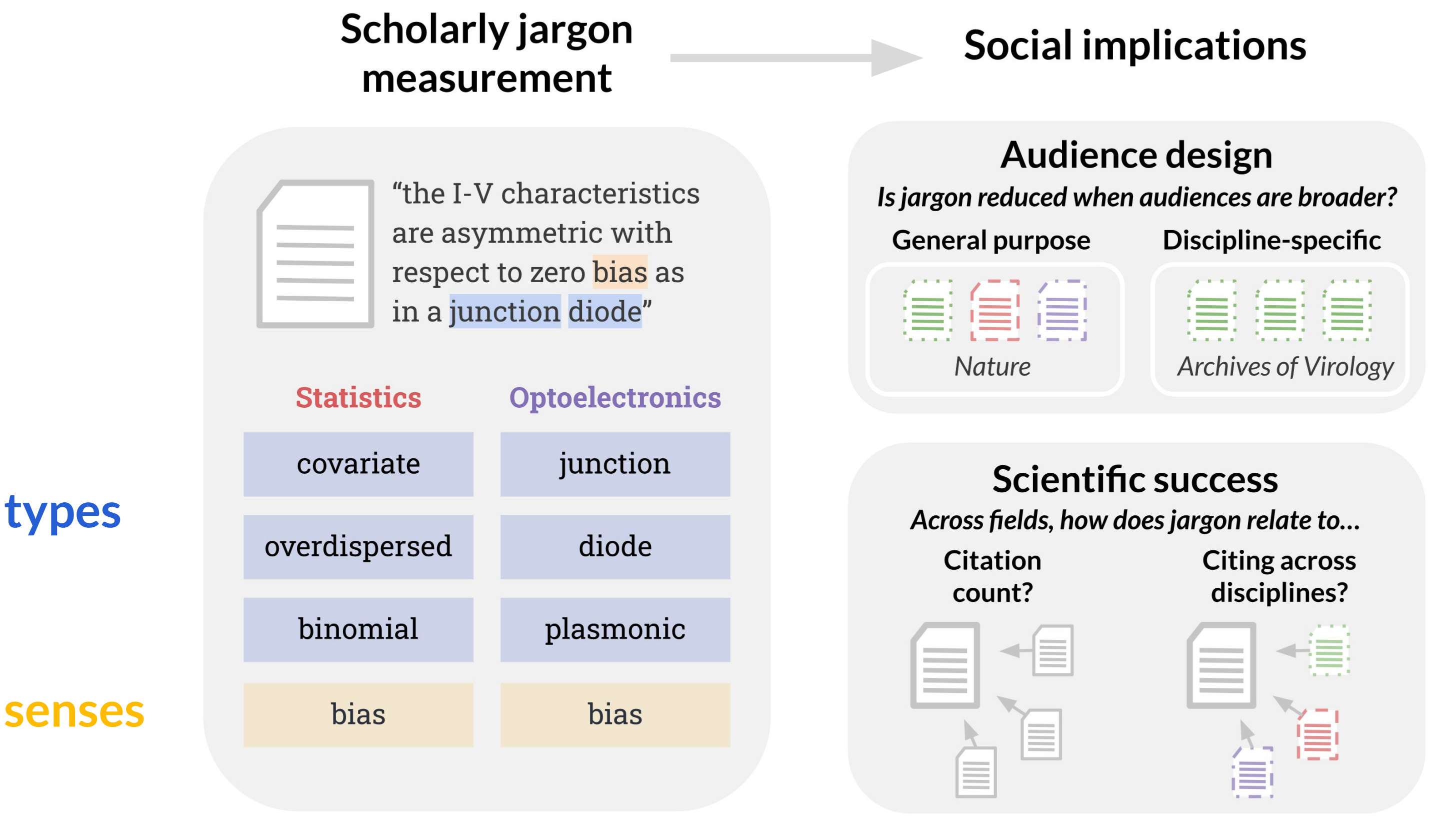


NPMI metric	AUC, recall
$\mathcal{S}_f(t)$, $\gamma = 0.5$	0.0550
$\mathcal{S}_f(t)$, $\gamma = 1.0$	0.0583
$\mathcal{S}_f(t)$, $\gamma = 1.5$	0.0631
$\mathcal{S}_f(t)$, $\gamma = 2.0$	0.0670
$\mathcal{S}_f(t)$, $\gamma = 2.5$	0.0697
$\mathcal{S}_f(t)$, dynamic γ	0.0675
$\mathcal{T}_f(t)$ baseline	0.0434

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.

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.



Scholarly jargon measurement

"the I-V characteristics are asymmetric with respect to zero bias as in a junction diode"

	Statistics	Optoelectronics
types	covariate	junction
	overdispersed	diode
	binomial	plasmonic
senses	bias	bias

Social implications

Audience design

Is jargon reduced when audiences are broader?

General purpose

Nature

Discipline-specific

Archives of Virology

Scientific success

Across fields, how does jargon relate to…

Citation count?

Citing across disciplines?