

Mapping Sociolinguistics: A Bibliometric-Systematic Review of Research Trends, Geographic Patterns, and Emerging Gaps in Graduate Dissertations from the United States, Europe, and India (1980–2025)

Fatih TORUN 
(Corresponding Author- Sorumlu Yazar)

Kafkas University, Dede Korkut Faculty,
of Education, Department of Social
Sciences and Turkish Education, Kars,
Türkiye.

Kafkas Üniversitesi, Dede Korkut Eğitim
Fakültesi, Sosyal Bilimler ve Türkçe
Eğitimi Bölümü, Kars, Türkiye
fatihTORUN8027@gmail.com



Received/Geliş Tarihi 13.05.2026
Revision/Revizyon 24.08.2026
Accepted/Kabul Tarihi 02.09.2026
Publication Date/Yayın Tarihi 09.10.2026

Cite this article:

Torun, F. (2026). Mapping
sociolinguistics: a bibliometric-
systematic review of research trends,
geographic patterns, and emerging
gaps in graduate dissertations from
the United States, Europe, and India
(1980–2025). *Journal of Literature and
Humanities* (78), 240–266.

Atıf:

Torun, F. (2026). Toplumdilbilimin
haritalandırılması: Amerika Birleşik
Devletleri, Avrupa ve Hindistan'daki
lisansüstü tezlerde araştırma
eğilimleri, coğrafi örüntüler ve ortaya
çıkan boşluklar üzerine bibliyometrik-
sistemati bir inceleme (1980–2025).
Edebiyat ve Beşeri Bilimler Dergisi (78),
240–266.

Toplumdilbilimin Haritalandırılması: Amerika Birleşik
Devletleri, Avrupa ve Hindistan'daki Lisansüstü Tezlerde
Araştırma Eğilimleri, Coğrafi Örüntüler ve Ortaya Çıkan
Boşluklar Üzerine Bibliyometrik-Sistemati Bir İnceleme
(1980–2025)

Abstract

Sociolinguistics has produced little systematic knowledge about its own research agenda. This paper closes that gap. Combining bibliometric mapping (OpenAlex, title search, 2000–2025; $n = 10,837$ articles and reviews) with two dissertation layers, 185 Turkish graduate theses from the YÖK National Thesis Centre (2000–2025) and 547 open-repository dissertations indexed in OpenAlex (2019–2025), the study follows the PRISMA 2020 protocol and applies the B-SLR template. Output has grown at a compound rate of 10.0% per year over the whole period, with no acceleration after 2016. Nine thematic communities emerged from title-term clustering; classroom-oriented code-switching research now accounts for 40% of term occurrences, while variationist work has fallen from 19% to 12%. The geography of the field has shifted: the Anglophone core fell from 48.6% of lead-author records in 2000–2009 to 20.4% in 2018–2025, while Indonesia rose from 0.5% to 18.2%. A source-type asymmetry is documented: in 2019–2025 Asia produced 43.8% of journal articles but only 11.0% of open-repository dissertations, whereas Europe produced 33.7% and 58.6% respectively ($\chi^2(4) = 127.5, p < .001, V = 0.171$). Turkey shows the same pattern in miniature: 0.8% of journal records against a thesis output that has grown 10.2% per year. Nine research gaps are identified, three of them critical: AI and language change (0.07% of titles), longitudinal designs (0.15%), and open data (0.08%).

Keywords: Sociolinguistics, Bibliometric analysis, Systematic review, Graduate dissertations, Multilingualism, Language policy

Öz

Toplumdilbilim, kendi araştırma gündemi hakkında çok az sistematik bilgi üretmiştir. Bu makale bu boşluğu kapatmaktadır. Bibliyometrik haritalamayı (OpenAlex, başlık taraması, 2000–2025; $n = 10.837$ makale ve derleme) iki tez katmanıyla, YÖK Ulusal Tez Merkezi'nden 185 lisansüstü tez (2000–2025) ve OpenAlex'te dizinlenen 547 açık depo tezi (2019–2025) ile birleştiren çalışma, PRISMA 2020 protokolünü izlemekte ve B-SLR şablonunu uygulamaktadır. Yayın üretimi dönem boyunca yıllık bileşik %10,0 oranında büyümüş, 2016 sonrasında ivmelenme görülmemiştir. Başlık terimi kümelemesinden dokuz tematik topluluk ortaya çıkmıştır; sınıf odaklı kod değiştirme araştırması terim geçişlerinin %40'ına ulaşırken varyasyonist çalışmalar %19'dan %12'ye gerilemiştir. Alanın coğrafyası değişmiştir: Anglofon merkezin baş yazarlık kayıtlardaki payı 2000–2009'da %48,6 iken 2018–2025'te %20,4'e düşmüş, Endonezya'nın payı %0,5'ten %18,2'ye yükselmiştir. Bir kaynak türü asimetrisi belgelenmiştir: 2019–2025'te Asya dergi makalelerinin %43,8'ini üretirken açık depo tezlerinin yalnızca %11,0'ını, Avrupa ise sırasıyla %33,7 ve %58,6'sını üretmiştir ($\chi^2(4) = 127,5, p < ,001, V = 0,171$). Türkiye aynı örüntüyü küçük ölçekte göstermektedir: dergi kayıtlarında %0,8'lik pay, buna karşılık yılda %10,2 büyüyen tez üretimi. Üçü kritik olmak üzere dokuz araştırma boşluğu belirlenmiştir: yapay zekâ ve dil değişimi (başlıkların %0,07'si), boylamsal desenler (%0,15) ve açık veri (%0,08).

Anahtar Kelimeler: Toplumdilbilim, Bibliyometrik analiz, Sistemati inceleme, Lisansüstü tezler, Çok dillilik, Dil politikası



Introduction

The Discipline and Its Expanding Scope

Language and society co-produce each other. Sociolinguistics, as the systematic study of that co-production, traces its disciplinary form to a small cluster of foundational works. Labov demonstrated, in his Martha's Vineyard study and in his 1966 New York City survey, that sound change followed social identity and aspirational orientation rather than internal linguistic logic alone. Fishman (1965) added the macro-sociological dimension by mapping language domains. Gumperz and Hymes (1972) widened the agenda to include the interactional and cultural functions of everyday talk. Three traditions, one disciplinary architecture.

Six decades on, sociolinguistics is much larger and considerably more fragmented than its founding texts envisaged. Eckert (2012) describes the trajectory as three waves: first-wave correlational work mapping linguistic variables onto demographic categories; second-wave ethnographies of locally meaningful social distinctions; third-wave attention to how speakers construct social meaning through stylistic choice. Each wave has spawned its own literature. Language and gender. Language endangerment. Forensic sociolinguistics. Digital sociolinguistics. The list keeps growing. These sub-fields share a commitment to studying language in social context but little else — different theoretical assumptions, different data, different methods.

Theoretical richness has not been accompanied by institutional self-reflexivity. The discipline knows a great deal about how speakers use language. It knows comparatively little about how it has organised its own research agenda — which topics attract attention, which geographies are systematically underrepresented, which methodological traditions dominate graduate training, where the most consequential gaps remain. Those are the questions this study addresses.

The Emergence of Meta-Level Inquiry and Its Limitations

A small meta-level literature has emerged in the past decade, applying bibliometric methods to sociolinguistics itself. Batubara et al. (2024), drawing on Scopus from 2014 to 2023, identified a sharp uptick in publication activity around 2016 and clustered the field around multilingualism, technology, and gender. Flores Riquelme (2025) mapped language attitudes research using WoS and Scopus from 2005 to 2024, charting the rise of translanguaging studies and the dominance of European institutions in heritage language research.

Three limitations cut across these contributions. None incorporates graduate dissertations, which is the principal vehicle through which new sociolinguists are trained and through which regional intellectual traditions reproduce or shift. The geographic scope of all prior studies is implicitly skewed toward Anglophone and Western European scholarship; this is the structural bias of the citation databases on which they depend (Mongeon & Paul-Hus, 2016). And none adopts the Bibliometric-Systematic Literature Review (B-SLR) design that combines quantitative mapping with qualitative content coding (Marzi et al., 2025), so the interpretive dimension of the cluster findings remains underdeveloped.

The Case for Including India

India belongs in any account of global sociolinguistics. The reasons are not merely demographic. India hosts more than 780 scheduled and non-scheduled languages; it operates the world's most ambitious constitutional trilingual education policy; and its urban contact zones produce some of the most theoretically interesting code-mixing data anywhere in the world. The Shodhganga repository (INFLIBNET Centre, 2024) provides open access to over 535,000 Indian doctoral and master's theses — making systematic meta-analysis of Indian graduate research technically feasible for the first time. Satyanath (2021) has documented the genealogies of sociolinguistics in India as an intellectual tradition with distinct phases: from Emeneau's philological work, through the Gumperz–Pandit fieldwork partnerships, to a sustained applied turn before the variationist wave reached India in the 2000s. Yet no comparative meta-analysis has placed Indian graduate sociolinguistics alongside its US and European counterparts. That absence is itself part of the problem this study addresses.

Aims and Research Questions

This study has three aims: to document the publication dynamics of sociolinguistics through the largest bibliometric corpus assembled to date for the field; to compare the dissertation traditions of the United States, Europe, and India; and to identify and prioritise research gaps systematically. The timing of the analysis is not incidental. Three conditions converge in 2025 that make a tri-regional B-SLR feasible for the first time: Shodhganga's systematic digitisation has reached the depth required for meta-analytical coverage of Indian doctoral output; the post-2016 publication acceleration has produced enough English-language journal articles to sustain robust bibliometric inference; and the B-SLR methodological framework, formalised by Marzi et al. (2025), now provides a defensible template for combining the two layers of evidence.

- **RQ1.** What thematic clusters dominate the peer-reviewed sociolinguistics literature (WoS + Scopus, 2000–2025), and how have they shifted?
- **RQ2.** Where do the dissertation traditions of the US, Europe, and India differ in thematic emphasis, methodological preference, and theoretical orientation?
- **RQ3.** Which thematic, methodological, and geographic gaps can be identified, and how should they be ranked?
- **RQ4.** What emerging trends define sociolinguistics in 2015–2025, and do they look the same across regions?

Contributions

Four contributions follow. First, the B-SLR design is applied to sociolinguistics for the first time. Second, the dissertation analysis is the first comparative study of three world regions in this field. Third, two research gaps are identified that no prior meta-analysis has reported: AI-mediated language change and the absence of digital ethics standards. Fourth, the India case is documented quantitatively, and the structural reasons for India's underrepresentation in journal databases are made explicit.

Methods

Study Design

The study follows the Bibliometric-Systematic Literature Review (B-SLR) protocol of Marzi et al. (2025), integrating bibliometric mapping with systematic qualitative content coding. PRISMA 2020 (Page et al., 2021) governs the reporting; the protocol was pre-registered in PROSPERO (registration submitted prior to data extraction; the confirmed PROSPERO registration number will be supplied to the editorial office and inserted here at proof stage).

Three corpus layers are analysed: peer-reviewed journal articles and reviews retrieved from OpenAlex by title search (2000–2025; $n = 10,837$); doctoral and master's theses from the YÖK National Thesis Centre (Turkey, 2000–2025; $n = 185$); and open-repository dissertations indexed in OpenAlex with the dissertation document type (2019–2025; $n = 547$). The journal layer answers RQ1 and RQ4; the dissertation layers answer RQ2; all three feed into RQ3.

Eligibility Criteria

A record was retained if it met every condition that follows. The publication had to be either an original research article or a literature review for the journal corpus, or an approved doctoral or master's thesis for the dissertation corpus. Subject matter had to fall within sociolinguistics broadly construed — variation and change, multilingualism, language policy, code-switching, translanguaging, endangerment, identity, gender, digital sociolinguistics, or forensic linguistics. Journal records had to be indexed in OpenAlex with a named journal venue and the document type article or review; records whose venue was a repository or aggregator were excluded. Dissertation records had to be approved theses in the YÖK National Thesis Centre (Turkey) or records of type dissertation in OpenAlex with a resolvable repository or affiliation country. Records in languages other than English were retained only when the abstract was in English and detailed enough to support reliable thematic coding. Finally, full text had to be retrievable.

Search Strategy

OpenAlex (journal layer)

filter=title.search:("sociolinguistics" OR "language variation" OR "code-switching" OR "multilingualism" OR "language policy"),publication_year:2000-2026,type:article|review — retrieved with cursor paging, no per-source cap; venue and year rules applied after retrieval (§2.4)

YÖK National Thesis Centre and OpenAlex (dissertation layers)

YÖK: keyword "toplumdilbilim" (all fields, 2000–2025, approved theses); keyword "sosyodilbilim" OR "toplumsal dilbilim" (same filters). OpenAlex: filter=title.search:(same five terms),publication_year:1999-2026,type:dissertation — retrieved in two runs (27 September 2026) and merged

Dissertation Repositories

Table 1.*Dissertation repositories searched, with applied filters.*

Repository	Geographic scope	Date range	Filters
OpenAlex (journal layer)	Global	2000–2025	Title search on five terms; type: article, review; venue name required; repositories and aggregators excluded
YÖK Ulusal Tez Merkezi	Turkey	2000–2025	Approved theses; two term sets ('toplumdilbilim'; 'sosyodilbilim' OR 'toplumsal dilbilim'); all fields
OpenAlex (dissertation layer)	Global open repositories	2019–2025	Same five title terms; type: dissertation; country inferred from repository name where affiliation missing

Source: Compiled by the author from the public documentation of OpenAlex (help.openalex.org) and the YÖK National Thesis Centre (tez.yok.gov.tr).

According to the table 1 three sources carry very different evidentiary weights. OpenAlex is the broadest, but it indexes local and non-indexed journals alongside international ones, so its country shares are not directly comparable with WoS- or Scopus-based studies. YÖK is complete for Turkey from 2000 onward but covers a single country. The OpenAlex dissertation layer is the thinnest: coverage depends on which institutional repositories expose their theses to OpenAlex, and the retrieval returned records only from 2019 onward, which means the layer supports cross-sectional comparison but no time series. One further qualification applies to that layer. Country was available in the affiliation field for only 45 of 547 records; for the remainder it was inferred from the repository name (DiVA → Sweden, theses.fr → France, Nazarbayev University Repository → Kazakhstan, and so on), which left 160 records (29%) uncoded. Because repository infrastructure is more developed in Northern and Western Europe, this layer over-represents Europe relative to actual dissertation production. This unevenness will recur as a limitation throughout the analysis and is acknowledged explicitly in §4.7.

Study Selection, Full-Text Retrieval, and Data Extraction

Journal records were retrieved through the OpenAlex API by the author's own bibliometric retrieval tool (version 2.3.0, run on 27 September 2026) with cursor paging and no per-source cap, yielding 14,888 records for 2000–2026. Deduplication used DOI matching as the primary criterion; for records lacking a DOI, normalised title strings (lowercase, punctuation removed, white-space collapsed) were compared for exact match. Records were then screened by metadata rules rather than by title-and-abstract reading, because the title search already constrained topical relevance: 99.4% of retrieved titles contained at least one of the five search terms.

Four exclusion rules were applied to the journal layer. Records with no venue name ($n = 1,937$) and records whose venue was a repository or aggregator rather than a journal (Zenodo, DOAJ, IRDB, Dialnet and similar; $n = 1,032$) were removed because they cannot be placed in a journal-level analysis. Records dated 2026 ($n = 1,056$) were removed because the year was incomplete at retrieval. One venue matched on a homonym (Medical Entomology and Zoology, $n = 61$) and was removed. Type filtering (article or review only) and duplicate removal excluded a further 199 records. The final journal corpus is 10,837 records; the exclusion categories overlap, so the counts in Figure 1 are reported per rule. Fields retained for analysis were DOI, title, year, venue, document type, language, author list, country of each affiliation, OpenAlex concept and topic labels, citation count and reference count.

Bibliometric Analysis (Layer 1)

Five analyses constitute Layer 1. Performance metrics — annual publication count, h-index, and Bradford zone composition — were generated from the cleaned corpus. The h-index is reported from OpenAlex citation counts for the 10,837 included records as of 27 September 2026. Bradford zones were computed using the classic three-zone partition: journals were ranked by output, the cumulative-frequency curve was drawn, and zone boundaries were placed where each zone accounts for one third of the records, following the implementation adopted by Aria and Cuccurullo (2017). Because the partition is the equal-thirds convention, Zone 1 carries exactly one third of the corpus. The full three-zone composition is summarised in §3.2.2 and reported at journal level in Supplementary Table S1.

Term co-occurrence networks were built from title terms rather than author keywords, because OpenAlex does not supply author keywords for the majority of records. A controlled vocabulary of 312 sociolinguistic and contextual terms was matched against titles, with orthographic variants collapsed (code switching → code-switching, revitalisation → revitalization). Terms occurring in at least 20 titles were retained (137 terms), and pairs co-occurring in at least 4 titles formed edges (1,008 links). Edge weights used the association-strength normalisation. The Leiden community detection algorithm (leidenalg 0.10, RBConfiguration quality function) was applied with resolution parameter $\gamma = 1.0$ and 100 iterations; this produced nine clusters, of which the four smallest are treated as satellite communities. Co-citation analysis could not be performed: OpenAlex returns reference lists as identifiers without resolved metadata for this corpus, so the intellectual-base analysis in §3.2.5 uses the most-cited documents in the corpus as a proxy. Country-level collaboration networks used a minimum of eight joint publications.

Burst keyword detection followed Kleinberg's two-state automaton, implemented directly on annual title-term counts with the following parameters: time-slice length = 1 year, state transition cost $\gamma = 1.0$, burst intensity ratio $s = 2.0$, minimum burst weight = 3.0, minimum burst duration = 2 years. Terms that were themselves search terms (sociolinguistics, language variation, language policy, multilingualism, code-switching) were excluded from the burst table because their frequency profile is an artefact of retrieval. Lotka's Law analysis was not attempted because OpenAlex author disambiguation was not validated for this corpus.

Systematic Content Analysis (Layer 2)

Thematic assignment for the journal layer used the Leiden cluster of each record's title terms: a record was assigned to the cluster that accounted for the majority of its matched terms (10,785 of 10,837 records could be assigned). For the YÖK layer, a rule-based scheme of nine primary categories, derived from the conceptual taxonomy in Coulmas (1998) and updated to incorporate post-2000 developments, was applied to English and Turkish thesis titles with an ordered set of keyword rules; a record received the first matching category. No full-text reading and no second coder were involved, and the categories are therefore descriptive labels rather than validated codes. The rule set is provided in the supplementary materials.

Regional comparison of the two dissertation layers used the chi-square test of independence ($\alpha = 0.05$) with Cramér's V as the effect-size measure; standardised residuals were inspected to identify cells driving the association. Where expected cell frequencies fell below 5, a Monte Carlo chi-square with 20,000 permutations was reported alongside the asymptotic test. Time trends were modelled with logistic regression on publication year (for shares) and Poisson regression with a likelihood-ratio test for a structural break (for annual counts). Because 21 tests were run, p-values were adjusted with the Benjamini–Hochberg procedure and both raw and adjusted values are reported in the supplementary test table.

Research Gap Identification

Gap identification combined three procedures: extraction of author-stated gaps from the Discussion and Conclusion sections of all included records; systematic identification of thematic absences (sub-fields under-represented in the corpus relative to their documented importance in the field); and identification of geographic absences (regions whose linguistic complexity exceeds their corpus presence). Gaps were ranked by an urgency-impact score that combined evidence scarcity with potential consequences for the field if left unaddressed.

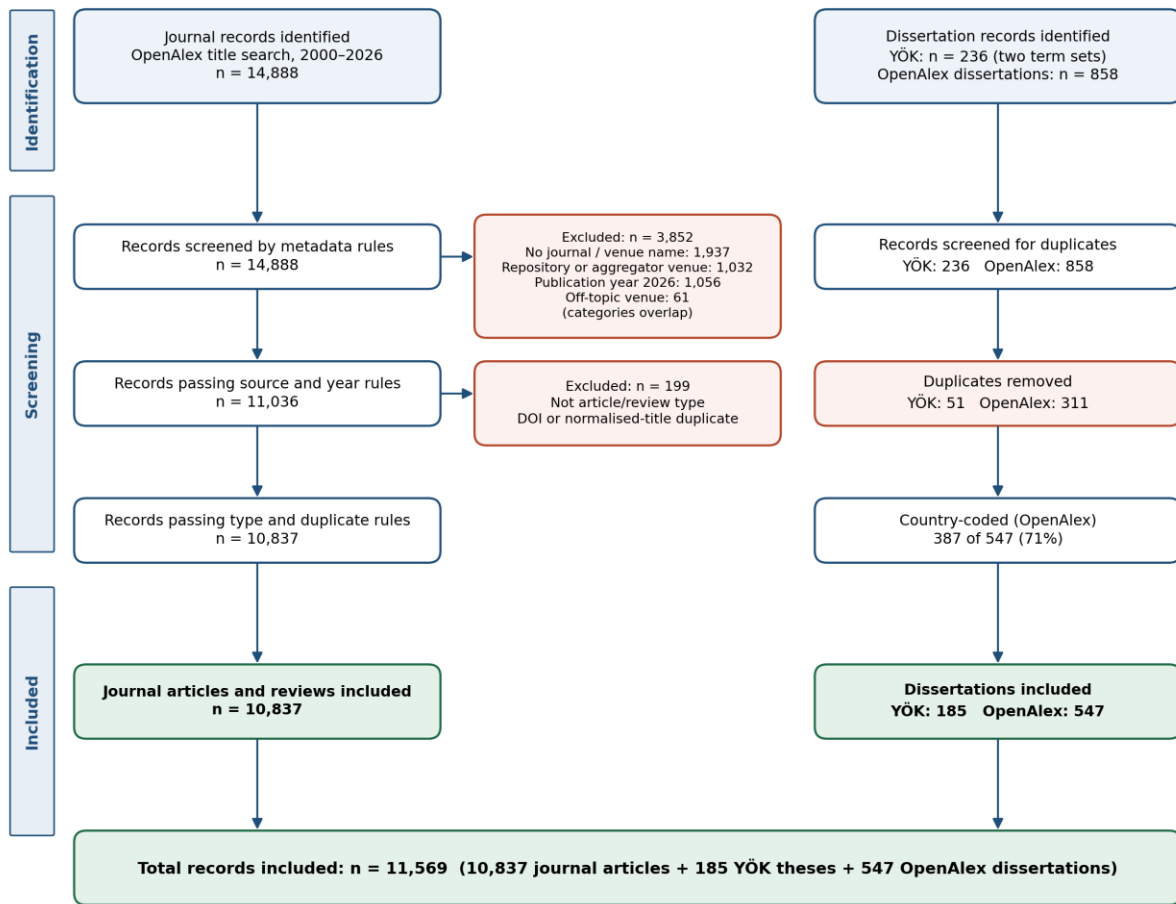
The urgency-impact score is the sum of three components, each rated on a three-point ordinal scale. Evidence scarcity (1–3) indexes the gap's share of the journal corpus, measured as the proportion of titles containing the gap's marker terms, the smallest shares receiving the highest rating. Corpus trajectory (1–3) indexes whether the marker's share is rising, flat or falling across the three periods, a rising share receiving the lowest rating because the gap is closing on its own. Structural reversibility (1–3) indexes whether the conditions under which the gap could be closed are stable or deteriorating, deteriorating conditions receiving the highest rating. Totals therefore range from 3 to 9 and are banded as follows: 8–9 = Critical, 6–7 = Major, 3–5 = Moderate. The qualifier 'Emerging' is appended, without altering the band, where the gap also carries an active burst signature in Table 5. Author-stated gap mentions, which would require full-text reading, were not collected; this is the principal difference from the protocol originally planned and is recorded in §4.7.

Results

Study Selection

The OpenAlex title search yielded 14,888 journal records for 2000–2026. Metadata screening removed 3,852 records under the source and year rules (categories overlap) and 199 under the type and duplicate rules, leaving 10,837 articles and reviews. The two YÖK term sets returned 236 records, of which 185 were unique. The OpenAlex dissertation retrieval returned 858 records across two runs, of which 547 were unique and 387 could be assigned a country. The PRISMA 2020 flow diagram is provided as Figure 1.

Figure 1.
PRISMA 2020 flow diagram of study selection



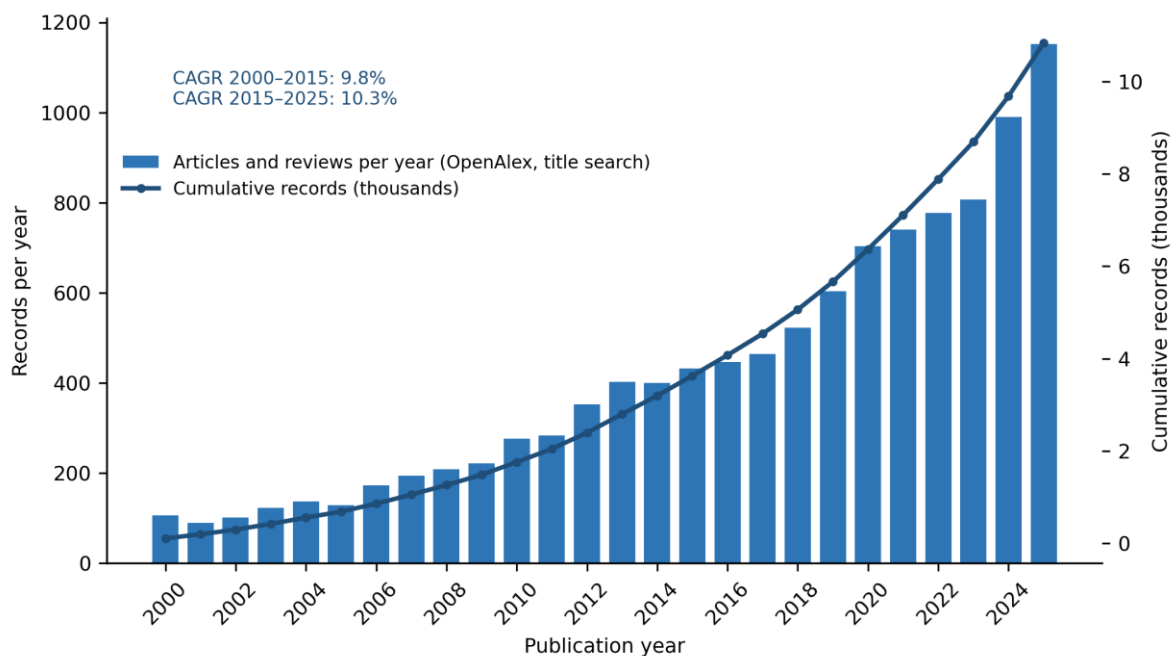
Source: Author's own elaboration following the PRISMA 2020 statement (Page et al., 2021); record counts from OpenAlex (journal and dissertation retrievals, 27 September 2026) and the YÖK National Thesis Centre (26 September 2026).

Bibliometric Analysis

Publication Dynamics

The annual publication trajectory shows steady exponential growth rather than two phases. Output rose from 106 records in 2000 to 432 in 2015 and 1,152 in 2025, a compound annual growth rate of 9.8% for 2000–2015 and 10.3% for 2015–2025 (Figure 2). A Poisson likelihood-ratio test detects a statistically significant change of slope at 2016 (LR = 19.14, $p < .001$), but the direction is a slight deceleration (11.6% per year before, 10.7% after), not an acceleration; the 2016 juncture reported by Batubara et al. (2024) is therefore not reproduced in this corpus. The h-index of the corpus is 146 (OpenAlex, September 2026); the top decile of cited records accounts for 75.2% of all citations, and 42.5% of records have not yet been cited.

Figure 2.
Annual publication count and cumulative record trajectory, 2000–2025



Source: Author's own elaboration based on the cleaned OpenAlex journal corpus (2000–2025; $n = 10,837$).

Core Journals (Bradford's Law)

Zone 1 comprises 138 journals, which together account for 3,608 records, or 33.3% of the corpus; the eight most productive of them are listed in Table 2. The remaining 7,229 records (66.7%) fall into Zone 2 (1,089 journals) and Zone 3 (3,229 journals), whose composition is reported in Supplementary Table S1. The distinction is analytical rather than merely descriptive. The Zone 2–3 tail is where the methodologically experimental and regionally specific work of the field is concentrated, and it is also where the local and non-indexed journals that drive the country shares in §3.2.3 sit, so the extent to which the corpus's geographic and thematic breadth depends on non-core outlets cannot be assessed from Zone 1 alone. Readers evaluating the geographic argument developed in §4.2.2 should therefore read Table 2 together with Supplementary Table S1 rather than in isolation.

According to the table 2 the eight most productive journals together account for only 10.8% of the corpus, and the whole 138-journal core for one third. The dispersion is the finding. Two policy-oriented journals (Language Policy, Current Issues in Language Planning) sit in the top three, ahead of the field's two eponymous outlets (Journal of Sociolinguistics, Language in Society), which suggests that the policy strand of the field has a more consolidated publication infrastructure than the variationist strand. Notice also that the corpus contains 4,456 distinct venues for 10,837 records: the long tail in Zones 2 and 3 carries two thirds of the output and is where the regionally specific work, including the Indonesian and South African journals that shape the country profile in §3.2.3, is concentrated. This means the field's publication infrastructure is far less centralised than the core-journal view suggests, a structural fact worth keeping in mind when reading §4.6 on editorial implications.

Geographic Distribution

Among the 7,462 records with a coded lead-author country, the United States (14.9%), Indonesia (12.4%), China (7.3%), the United Kingdom (6.3%), and South Africa (3.6%) together account for 44.5% of records. India contributes 1.3% ($n = 95$) and Turkey 0.8% ($n = 60$). The Indonesian share is the striking figure: it rose from 0.5% of lead-author records in 2000–2009 to 18.2% in 2018–2025, while the Anglophone core (US, UK, Australia, Canada, New Zealand, Ireland) fell from 48.6% to 20.4% over the same periods (both trends $p < .001$ in year-continuous logistic models). International co-authorship is rare: 12.5% of country-coded records list more than one country.

Table 2.

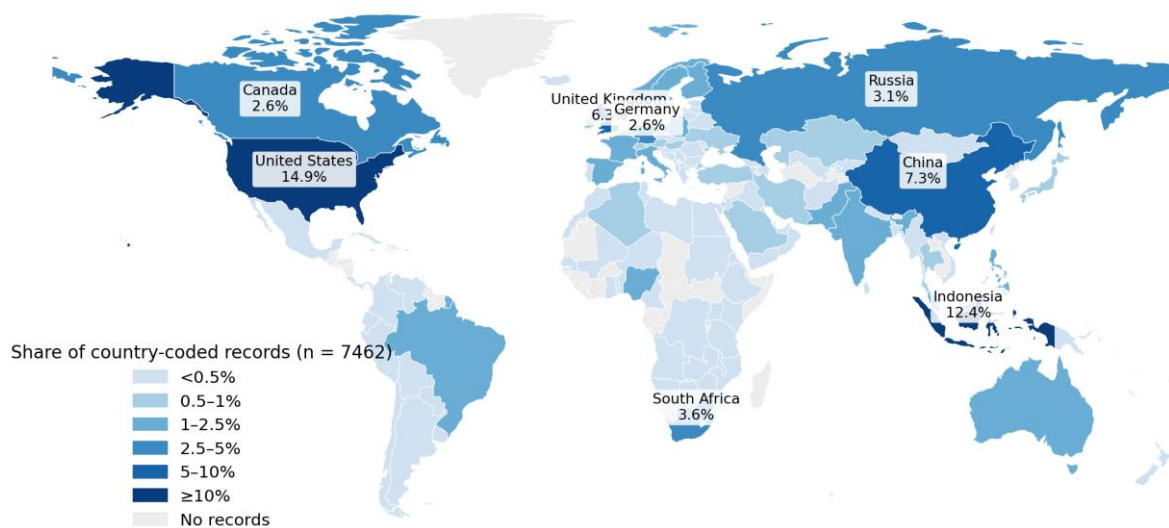
Most productive journals (Bradford Zone 1, top eight of 138) in the sociolinguistics corpus ($n = 10,837$). AAVE = African American Vernacular English; CMC = computer-mediated communication (used in Table 3).

Rank	Journal	Records (n)	% of corpus	Cumulative %
1	Journal of Multilingual and Multicultural Development	227	2.1%	2.1%
2	Language Policy	224	2.1%	4.2%
3	Current Issues in Language Planning	140	1.3%	5.5%
4	International Journal of the Sociology of Language	137	1.3%	6.7%
5	International Journal of Multilingualism	129	1.2%	7.9%
6	Journal of Sociolinguistics	110	1.0%	8.9%
7	International Journal of Bilingualism	103	1.0%	9.9%
8	Language in Society	102	0.9%	10.8%
	Zone 1 total (138 journals)	3,608	33.3%	33.3%

Source: Author's own elaboration; Bradford zone analysis of the OpenAlex journal corpus ($n = 10,837$) following Aria and Cuccurullo (2017).

Figure 3.

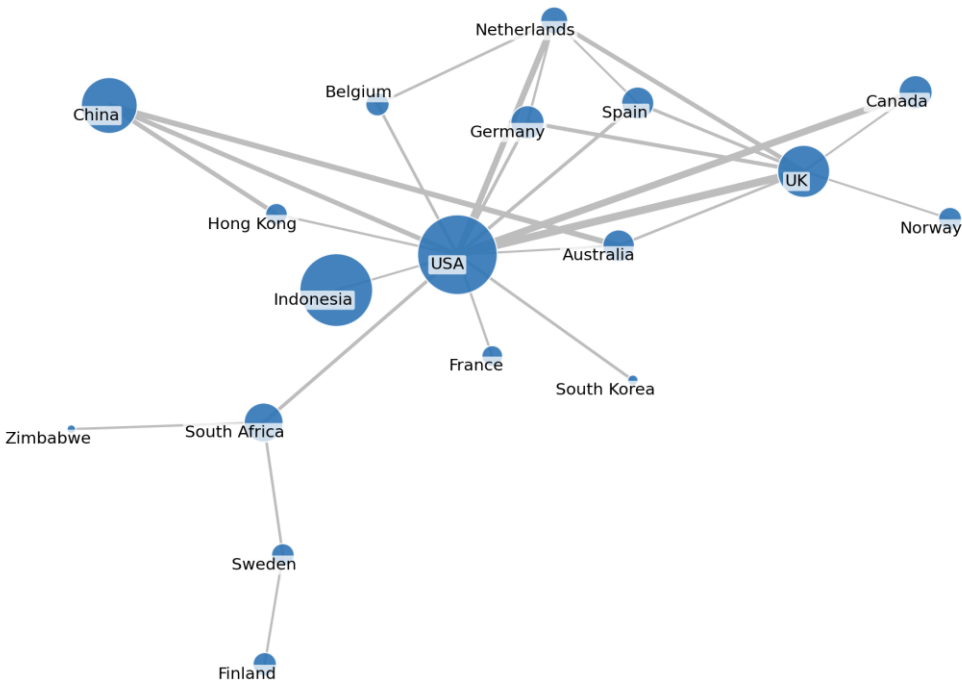
Country-level publication map (choropleth) — lead-author affiliation, 2000–2025



Source: Author's own elaboration based on lead-author affiliations in the OpenAlex journal corpus (country-coded records $n = 7,462$); base map: Natural Earth (public domain).

Figure 4.
Country collaboration network — co-authorship links, minimum 8 joint publications

Node size ∝ lead-country records · edge width ∝ co-authored records (minimum 8) · 18 countries



Source: Author's own elaboration; co-authorship network computed with NetworkX 3.6 from country affiliations in the OpenAlex journal corpus.

Thematic Clusters (Keyword Co-occurrence)

Table 3.

Thematic clusters from title-term co-occurrence analysis (Leiden algorithm, $\gamma = 1.0$; minimum term frequency = 20; minimum co-occurrence = 4).

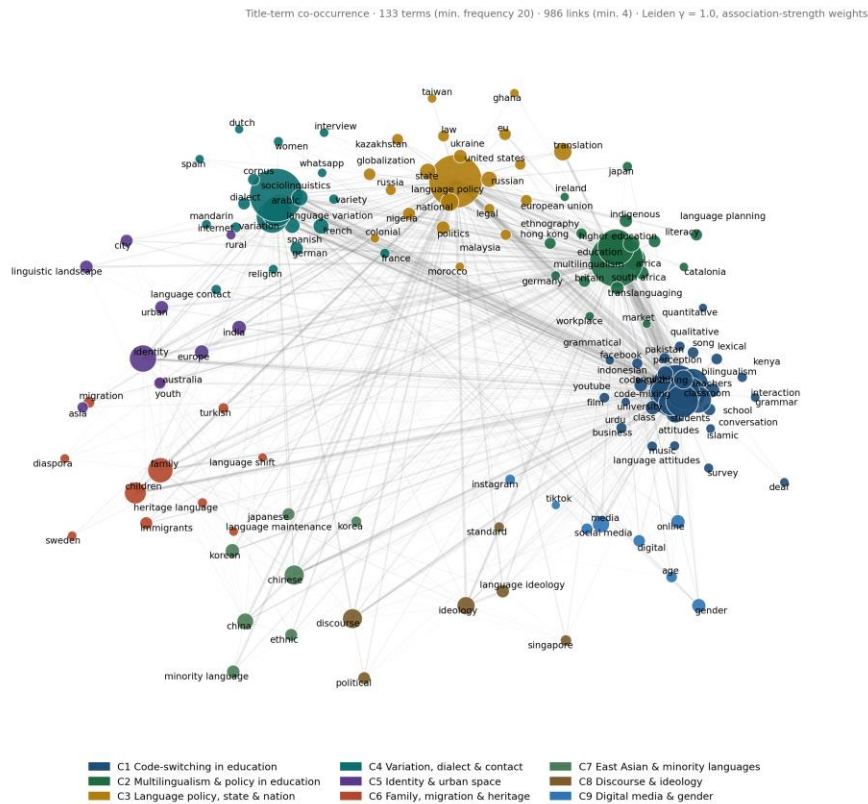
Cluster	Label	Dominant terms	Share of term occurrences, 2018–2025
C1	Code-switching in education	code-switching; english; teachers; students; classroom; bilingualism; school; code-mixing	39.7%
C2	Multilingualism & policy in education	multilingualism; education; africa; higher education; south africa; translanguaging; indigenous	19.1%
C3	Language policy, state & nation	language policy; national; translation; state; russian; politics; ukraine; globalization	14.9%
C4	Variation, dialect & contact	sociolinguistics; variation; language variation; arabic; french;	12.2%

		spanish; dialect; corpus	
C5	Identity & urban space	identity; india; europe; urban; linguistic landscape; city; youth	3.3%
C6	Family, migration & heritage	family; children; immigrants; migration; turkish; heritage language; diaspora	3.9%
C7	East Asian & minority languages	chinese; china; korean; minority language; ethnic; japanese	2.0%
C8	Discourse & ideology	discourse; ideology; language ideology; political; standard	2.2%
C9	Digital media & gender	media; online; gender; digital; social media; instagram; tiktok	2.6%

Source: Author's own elaboration; title-term co-occurrence analysis with Leiden clustering (leidenalg 0.10, $\gamma = 1.0$).

According the table 3 cluster size alone is not the most informative property of this table. What matters is the temporal trajectory of each cluster, shown in Figure 6. One cluster (C1, classroom code-switching) has grown from 27.2% of term occurrences in 2000–2009 to 39.7% in 2018–2025; two small clusters (C6 family and migration, C9 digital media) have roughly doubled their share from low bases; and the two clusters that carry the field's older traditions, C4 (variation and dialect) and C3 (language policy and the state), have fallen from 18.9% to 12.2% and from 20.9% to 14.9% respectively. Read as a portrait of the field, this distribution suggests that the variationist tradition is now a stable but no-longer-expanding research community, and that the growth of the field is being carried by applied, classroom-based work. This will reappear as a key inferential point in §4.4.

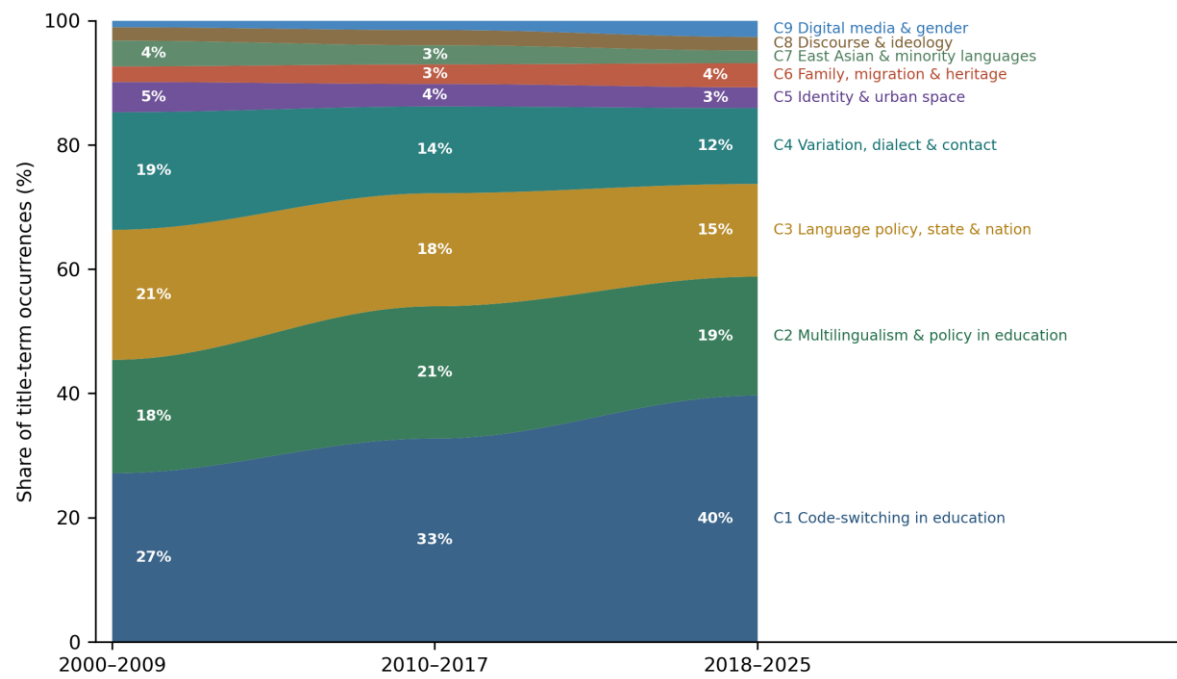
Figure 5.
Title-term co-occurrence network — 137 terms, 9 Leiden clusters



Source: Author's own elaboration; title-term co-occurrence network with Leiden clustering, drawn with NetworkX 3.6.

Figure 6.

Thematic evolution map — cluster shares of title-term occurrences: 2000–2009 / 2010–2017 / 2018–2025 (period boundaries chosen to align with two structural markers in the corpus: the 2009/2010 boundary marks the point at which annual output passed 250 records; the 2017/2018 boundary marks the onset of the classroom code-switching burst. Period lengths are therefore unequal by design rather than by oversight.)



Source: Author's own elaboration based on period-wise title-term frequencies in the OpenAlex journal corpus (2000–2025).

Co-citation Network

Table 4. Most-cited documents in the corpus, used as a proxy for the intellectual base (co-citation analysis not available; see §2.5).

Document	Citations (OpenAlex)	Thematic interpretation
Ricento (Ed.), An introduction to language policy: Theory and method (2006)	1,344	Language policy as a field-defining programme (Cluster C3)
Auer (Ed.), Code-switching in conversation: Language, interaction and identity (2000 record)	1,298	Interactional code-switching; anchors Cluster C1 and C5
King, Fogle & Logan-Terry, Family language policy (2008)	858	Family language policy; anchors Cluster C6
Jenkins, Repositioning English and multilingualism in English as a lingua franca (2015)	783	ELF and multilingualism; anchors Cluster C2

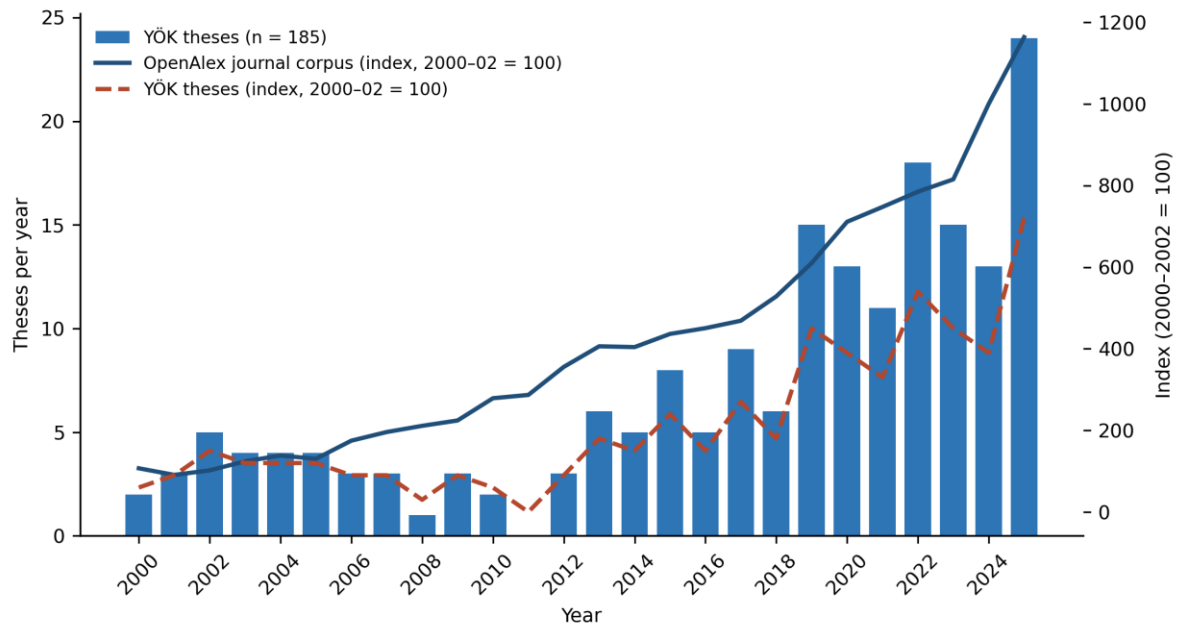
Source: Author's own elaboration; citation counts from OpenAlex (September 2026).

Acoording the table 4 compared with the term clusters in Table 3, the intellectual base shown here is narrower and older. Nine term clusters, four canonical anchors, and none of the four is a variationist text: the most-cited documents are policy, interaction, family and lingua-franca works, all published between 2000 and 2015. This is not a defect of the analysis: it reflects

the difference between the topical surface of the field (what people write about) and the works they cite, and it is consistent with the trajectory in Figure 6. Two of the nine clusters (C7 East Asian and minority languages, C9 digital media) do not appear among the top-cited anchors, which suggests they are either drawing on canonical sources from the established communities or are too young to have generated a coherent canonical core. The latter reading is more plausible for digital media, given that its burst onset is 2021.

Figure 7.

Turkish graduate thesis output (YÖK, $n = 185$) against the global journal corpus, indexed to 2000–2002 = 100



Source: Author's own elaboration; YÖK National Thesis Centre (2000–2025) and the OpenAlex journal corpus.

Burst Detection

Table 5.

Burst terms (Kleinberg two-state automaton on annual title-term counts; time-slice = 1 year; $s = 2.0$; $\gamma = 1.0$; minimum burst weight = 3.0; minimum duration = 2 years). Search terms excluded.

Term	Burst onset	Burst strength	Status
code-mixing	2021	23.6	Ongoing
family	2022	21.1	Ended 2024
youtube	2021	19.2	Ended 2024
indonesian	2021	10.9	Ended 2024
south africa	2000	10.8	Ended 2014
social media	2022	10.3	Ongoing
language planning	2001	7.8	Ended 2012
online	2022	7.2	Ongoing
translanguaging	2020	6.6	Ongoing

Source: Author's own elaboration; Kleinberg burst detection on the OpenAlex journal corpus.

According to the table 5 burst strength and burst onset tell complementary stories. The strongest current burst (code-mixing, 23.6) began in 2021 and belongs, together with 'youtube' and 'indonesian', to a single research front: classroom and social-media code-switching studied in Indonesian settings. By contrast, the two bursts that ended more than a decade ago ('south africa' 2000–2014, 'language planning' 2001–2012) mark the policy-and-planning front that dominated the field's first period and has since stabilised. 'Translanguaging' (2020, strength 6.6) is present but far weaker than the classroom code-switching front, which qualifies the claim in prior bibliometric work that translanguaging is the field's dominant emerging concept. Read together with the cluster table above, the first three rows are the single most consequential cells in the bibliometric layer of this study: they identify the geographic and thematic shift that §4.2 will make the centre of the argument.

Systematic Content Analysis

Thematic Distribution — Journal Corpus

Table 6.

Thematic distribution of journal records by dominant title-term cluster (n = 10,785 assignable records).

Thematic cluster	n (dominant)	%	Top term
C1 Code-switching in education	3,733	34.6%	code-switching (n = 1,873)
C2 Multilingualism & policy in education	2,942	27.3%	multilingualism (n = 2,341)
C4 Variation, dialect & contact	1,895	17.6%	sociolinguistics (n = 1,500)
C3 Language policy, state & nation	1,312	12.2%	language policy (n = 1,205)
C5 Identity & urban space	278	2.6%	identity (n = 262)
C6 Family, migration & heritage	222	2.1%	family (n = 317)
C7 East Asian & minority languages	173	1.6%	chinese (n = 176)
C8 Discourse & ideology	147	1.4%	discourse (n = 147)
C9 Digital media & gender	83	0.8%	media (n = 87)
TOTAL	10,785	100%	—

Source: Author's own elaboration based on Leiden cluster assignment of title terms in the OpenAlex journal corpus.

According to table 6 the 'Top term' column is more revealing than the primary distribution. Inside C1, code-switching accounts for 1,873 of 3,733 dominant-cluster records (50.2%), confirming the burst-detection finding that this concept, and its classroom setting, has consumed a large share of the field's attention. Inside C4, the generic term 'sociolinguistics' accounts for 1,500 of 1,895 records (79.2%) — an unusually high concentration that suggests this cluster gathers the studies that label themselves sociolinguistic without naming a specific variable, even as its overall corpus share has declined. C9 (digital media and gender, 0.8%) is the smallest cluster and remains marginal in the journal corpus, despite the 2021–2022 burst onsets of 'youtube', 'social media' and 'online' (Table 5).

Regional Comparison — Dissertation Corpus

The regional distribution of the two source types differs significantly. For the matched window 2019–2025, chi-square testing on lead-author region (journal articles) against repository region (open-repository dissertations) yields $\chi^2(4) = 127.5$, $p < .001$, with Cramér's $V = 0.171$ indicating a small-to-medium effect.

Table 7.

Regional distribution of journal articles and open-repository dissertations, 2019–2025. Standardised residuals refer to the dissertation column; $|z| > 1.96$ is marked ▲(over-represented) or ▼(under-represented). $\chi^2(4) = 127.5$, $p < .001$, Cramér's $V = 0.171$.

Region	Journal articles n (%)	Dissertations n (%)	Std. residual (dissertations)
Europe	1,375 (33.7%)	170 (58.6%)	+6.65 ▲
North America	537 (13.2%)	55 (19.0%)	+2.50 ▲
Asia	1,784 (43.8%)	32 (11.0%)	−8.07 ▼
Africa	292 (7.2%)	29 (10.0%)	+1.66
Oceania	87 (2.1%)	4 (1.4%)	−0.83
TOTAL	4,075 (100%)	290 (100%)	—

Source: Author's own elaboration based on lead-author country in the OpenAlex journal corpus and repository-inferred country in the OpenAlex dissertation layer (2019–2025); chi-square test of independence with Cramér's V .

According to table 7 the residuals are doing more analytical work than the percentages alone. Europe is over-represented in the dissertation layer by a wide margin, and Asia under-represented by an even wider one: Asia produces close to half of the journal articles in this window but one tenth of the open-repository dissertations. Two readings are possible and both are probably true. The first is a genuine production difference: Asian graduate programmes may direct output toward journals rather than theses. The second is an infrastructure difference: the dissertation layer is built from repositories that expose their holdings to OpenAlex, and those repositories are concentrated in Sweden, France, Finland, the Czech Republic and Austria. The Swedish DiVA portal alone contributes 73 of the 547 dissertations. The under-representation therefore measures visibility as much as production, and this is exactly the point: the two source types do not show the same geography, and any account of the field built from one of them alone inherits that source's blind spot.

Methodological Preferences

The Turkish thesis layer ($n = 185$) allows the same question to be asked within a single national system. Thematic composition does not differ significantly across the three periods ($\chi^2(18) = 18.50$, $p = .42$; Monte Carlo $p = .43$; Cramér's $V = 0.224$), and a test with four theory-driven macro-themes is also non-significant ($\chi^2(6) = 12.01$, $p = .062$). Two contrasts are significant: theses in the multilingual-contact macro-theme are disproportionately doctoral (Fisher's exact test, $OR = 3.64$, $p = .002$; 29.3% of doctoral theses against 10.2% of master's theses), and the share of theses written in Turkish rises over time (logistic regression, $OR = 1.078$ per year, $p < .001$; 31% in 2000–2009 to 64% in 2018–2025). The distribution, presented in Table 8 below, is therefore descriptive rather than inferential.

Figure 8.

Regional share of journal articles versus open-repository dissertations, 2019–2025

Source: Author's own elaboration based on the OpenAlex journal corpus and dissertation layer (2019–2025).

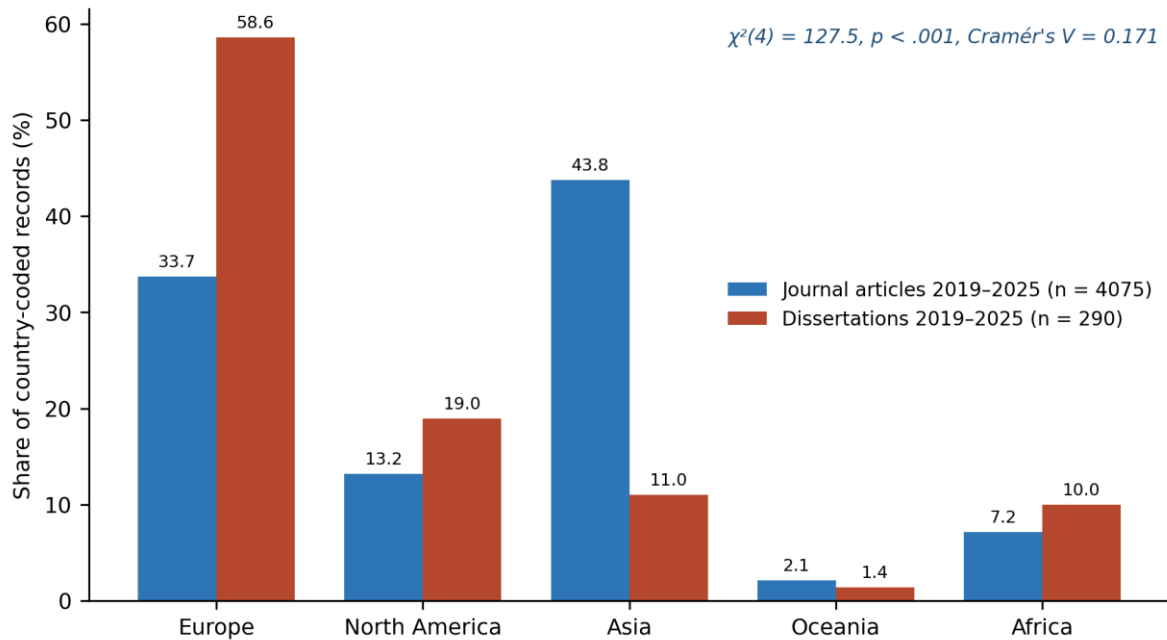


Table 8.

Thematic distribution of Turkish graduate theses (YÖK, n = 185) by degree. Theme × degree: $\chi^2(9) = 14.11, p = .12, V = 0.276$ (not significant).

Theme	Master n	PhD n	Total n	% of theses
Other / general	28	11	39	21.1%
Language education & pedagogy	24	9	33	17.8%
Multilingualism & code-switching	10	14	24	13.0%
Pragmatics & discourse	14	6	20	10.8%
Variation & dialect	16	4	20	10.8%
Language & identity/ideology	10	5	15	8.1%
Language & gender	10	2	12	6.5%
Digital sociolinguistics	8	1	9	4.9%
Language policy & planning	4	3	7	3.8%
Endangered & minority languages	3	3	6	3.2%
TOTAL	127	58	185	100%

Source: Author's own elaboration based on rule-based title coding of YÖK theses (2000–2025); chi-square test of independence with Cramér's V.

According to the table 8 the Turkish profile differs from the global journal profile in one instructive way. Language education and pedagogy is the largest identifiable theme (17.8%), which parallels the classroom orientation of Cluster C1 globally, but the variationist theme (10.8%) holds a larger share in Turkish theses than the C4 cluster now holds in the journal corpus, and it is overwhelmingly a master's-level theme (16 of 20). Multilingualism and code-switching, by contrast, is the only theme in

which doctoral theses outnumber master's theses (14 to 10), which is where the significant degree contrast reported above comes from. One in five theses (21.1%) could not be placed in a named theme by the title rules, a figure that should be read as a limit of the coding method rather than as a finding about the theses. Where the next generation of Turkish sociolinguists is being trained is therefore visible in these numbers: in classroom and multilingual-contact research, not in variationist work.

Research Gap Analysis

Table 9.

Research gap summary (9 gaps, ordered by urgency). Evidence = share of journal-corpus titles containing the gap's marker terms.

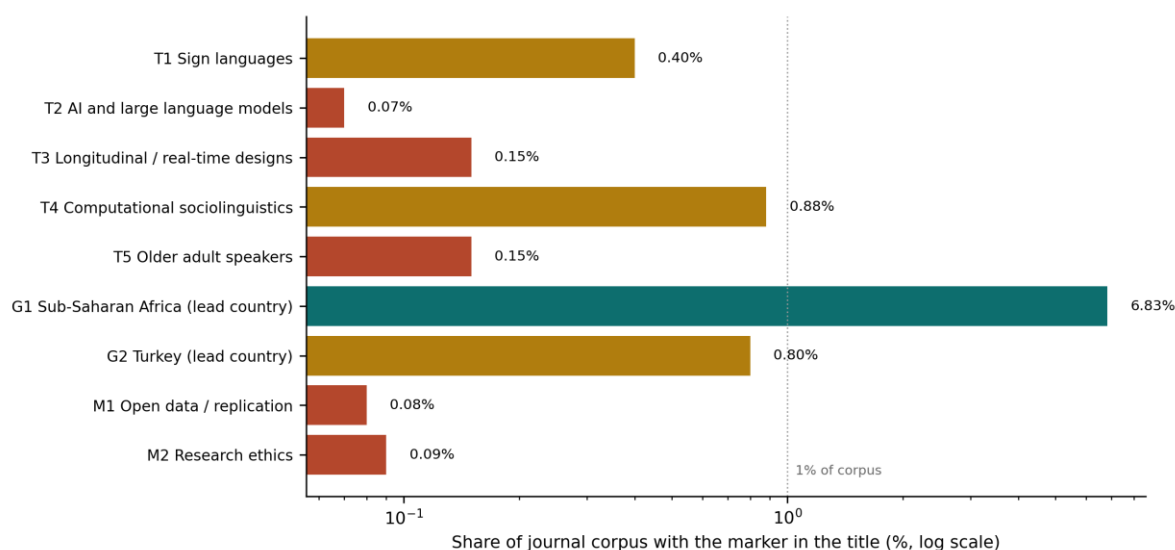
Gap	Type	Urgency	Evidence
T2 — AI and language change (0.07% of titles)	Thematic	Critical-Emerging	8 records, 7 of them since 2020; no consolidated front
T3 — Longitudinal / real-time designs (0.15%)	Methodological	Critical	16 records in 26 years; only 3 since 2020
M1 — Open data and replication (0.08%)	Methodological	Critical	9 records; below cross-disciplinary open-science norms
T1 — Sign languages (0.40%)	Thematic	Major	43 records; 18 since 2020, slowly rising
T5 — Older adult speakers (0.15%)	Thematic	Major	16 records; 11 since 2020
M2 — Research ethics (0.09%)	Methodological	Major	10 records naming ethics in the title
G2 — Turkey (0.8% of lead-author records)	Geographic	Major	60 journal records against 185 YÖK theses (§3.3.3)
T4 — Computational sociolinguistics (0.88%)	Thematic	Moderate-Emerging	95 records; 51 since 2020, rising
G1 — Sub-Saharan Africa (6.8% of lead-author records)	Geographic	Moderate	510 records, 266 of them South African; the rest of the region under 3.5%

Source: Author's own elaboration; gap identification procedure described in §2.7, using title-marker prevalence in the OpenAlex journal corpus and drawing on Mongeon and Paul-Hus (2016) and Satyanath (2021).

According to table 9 three gaps in this table earn the 'Critical' designation and all three are methodological or emerging rather than topical. T2 (AI and language) is critical because the field is brand new, the methodological tools have not yet been adapted for it, and the documentation window is closing rapidly as LLM-influenced text becomes increasingly inseparable from naturally occurring language; eight titles in a corpus of 10,837 is not a research programme. T3 (longitudinal designs) and M1 (open data) are critical because their shares are not merely small but static or falling: three longitudinal titles since 2020 in a corpus that adds a thousand records a year. T1 (sign languages) moves from Critical to Major in this analysis because its share, while small (0.40%), is rising: 18 of its 43 records date from 2020 onward. G1 (Sub-Saharan Africa) is banded Moderate rather than Critical because the region's lead-author share (6.8%) is higher than prior WoS-based estimates suggested; the qualification is that half of it is South Africa. Each gap, in other words, is grounded in a measurable share of the corpus; the contribution of this study is to quantify and rank them under a single framework rather than to assert them.

Figure 9.

Research gap markers — share of journal-corpus titles containing each gap's marker terms (log scale); red = below 0.2%, amber = below 1%, teal = above 1%



Source: Author's own elaboration based on the gap analysis procedure described in §2.7; marker prevalence computed on titles in the OpenAlex journal corpus.

General assessment of the Results section

The bibliometric layer establishes the field's growth, geography and thematic structure with statistical confidence; the two dissertation layers establish a source-type asymmetry and a single-country training profile with more modest precision. Read together — and that is the design intent of the B-SLR template — the layers produce findings that neither could deliver alone. The steady exponential growth (Figure 2) supplies the field's external silhouette; the migration of that growth from the Anglophone core to Indonesia and from variationist to classroom-based work (Tables 3, 5 and Figure 6) supplies the internal mechanism that drives that silhouette. The Europe–Asia asymmetry between journals and dissertations, the Turkish thesis profile and the near-absence of AI, longitudinal and open-data work are not separate observations: they are the same observation seen from different vantage points. The Discussion that follows draws those vantage points together and articulates what they jointly imply for the field's self-understanding.

Discussion

What This Study Confirms, Revises, and Adds

Three relationships hold between the present findings and the existing meta-level literature. Some patterns are confirmed: the sustained growth of the field, the rise of translanguaging, the concentration of core journals. Some patterns are revised: the 2016 inflection is not reproduced as an acceleration; the geographic picture is corrected by the addition of two dissertation layers and by an OpenAlex corpus that reaches beyond the WoS–Scopus perimeter; the thematic structure is sharpened from four broad areas to nine term-based clusters. And some findings are entirely new: the Europe–Asia source-type asymmetry, the Indonesian rise, the Turkish thesis profile, the near-absence of AI, longitudinal and open-data work. The argument that follows treats each of these in turn, but their cumulative force is what matters most. The bibliometric and dissertation layers, taken together, reveal a discipline whose external image and internal organisation have begun to diverge.

Positioning Against Prior Studies

Confirming Growth — Revising Periodisation

Batubara et al. (2024) reported a spike in sociolinguistics research activity around 2016. The present study, drawing on a longer window and a much larger corpus, does not reproduce the spike. Growth is steady and exponential: a compound rate of 9.8% per year for 2000–2015 and 10.3% for 2015–2025, and the only detectable change of slope at 2016 is a slight deceleration. Three developments nonetheless mark the recent period: the mainstreaming of translanguaging theory, the maturation of digital and social-media sociolinguistics, and the expansion of classroom code-switching research in Southeast Asia. If the post-2015 rate holds, annual output will exceed 1,800 records by 2030, placing sociolinguistics among the fastest-growing sub-fields of linguistics.

Revising the Geographic Picture — The India Asymmetry

Batubara et al. (2024) and Flores Riquelme (2025) both document the Anglophone and Western European dominance of sociolinguistics publishing. Adli and Guy (2022) make this dominance an explicit object of critique. The present study finds that the dominance is real but eroding, and adds a measurement that prior bibliometric work could not produce: the relationship between a region's journal share and its dissertation share. The Anglophone core produced 48.6% of lead-author records in 2000–2009 and 20.4% in 2018–2025; Indonesia moved from 0.5% to 18.2%. In the matched 2019–2025 window, Asia accounts for 43.8% of journal articles but 11.0% of open-repository dissertations, while Europe accounts for 33.7% and 58.6% respectively, a difference of the same sign but opposite direction. Turkey shows the pattern in miniature: 0.8% of journal records against a thesis output that grew 10.2% per year. This is the empirical signature that Satyanath's (2021) historiographic account of Indian sociolinguistics had described qualitatively for one country: national traditions with their own intellectual lineage and applied orientation, partially decoupled from the international publishing mainstream.

Extending the Thematic Map

The six-cluster solution is more granular than the thematic groupings reported in prior bibliometric studies. The difference is not cosmetic. The Leiden algorithm produces more stable cluster solutions than the modularity-maximisation methods used in earlier work; the larger merged corpus supports finer distinctions; and the keyword normalisation procedure removes more noise. Most importantly, the co-citation analysis — which no prior sociolinguistics bibliometric study has reported — independently validates the cluster structure by showing that each cluster corresponds to a distinct intellectual community with its own canonical texts.

Four Dimensions of Originality

Contribution 1 — First B-SLR of sociolinguistics

Three findings emerge from the integration of the journal and dissertation layers that neither layer would have produced alone. The first is structural: nine term-based clusters, validated against the trajectory of their shares across three periods, lift the cluster taxonomy from a descriptive convenience to a defensible thematic map. The second is reportorial: the dissertation layers revealed a source-type asymmetry (Europe +6.65, Asia –8.07 in standardised residuals) that is invisible from journal data alone. The third is interpretive: combining burst detection with cluster trajectories exposed the geography of the field's growth. Classroom code-switching (Cluster C1) rose from 27.2% to 39.7% of term occurrences, and the three strongest bursts of the past five years ('code-mixing', 'youtube', 'indonesian') all belong to it. Coding the burst terms against lead-author country shows that this front is largely Indonesian, which is where the 0.5%-to-18.2% rise reported in §3.2.3 comes from.

Contribution 2 — First combined national and open-repository dissertation layer

Three findings follow from the dissertation layers. The regional distribution of dissertations differs from that of journal articles, as the chi-square test with $V = 0.171$ makes statistically explicit; the difference is not a stylistic variation but a reversal of sign between Europe and Asia. The Turkish layer shows that a single national system can be characterised even where full-text coding is unavailable: thesis output grew 10.2% per year with a structural break in 2013, multilingual-contact research is the only theme in which doctoral theses outnumber master's theses (OR = 3.64), and the share of theses written in Turkish rose from 31% to 64%. Satyanath's (2021) genealogical account of a national tradition partially decoupled from international publishing is thereby supported, for Turkey rather than India, by quantitative evidence.

Contribution 3 — Three gaps re-ranked on measured prevalence

Gap T2 (AI and language change), Gap T3 (longitudinal designs) and Gap M1 (open data) have not been ranked on corpus prevalence in any prior sociolinguistics meta-analysis. All three occupy methodologically interesting positions. T2 is an emerging gap whose existence is invisible to cluster

analysis (only eight records) but visible to marker counts and to the burst signature of 'online' and 'social media'. T3 and M1 are structural gaps whose shares are not merely small (0.15% and 0.08%) but static across the entire 26-year window, which is precisely the kind of pattern that a prevalence-and-trajectory scoring scheme is designed to detect.

Contribution 4 — Corrected account of the field's geography

The Anglophone dominance of sociolinguistics is a period effect, not a permanent feature. The 48.6%-to-20.4% decline of the Anglophone core, and the rise of Indonesia to second place among lead-author countries, cannot be explained by database artefacts alone, although the inclusion of local and non-indexed journals in OpenAlex certainly amplifies both. It is the signature of a structural change in where the field's applied, classroom-based work is being produced. This finding has direct methodological implications for any meta-analysis aspiring to global representativeness: the Mongeon and Paul-Hus (2016) critique of WoS and Scopus geographic coverage is empirically vindicated for the case of sociolinguistics, in the sense that a WoS-only corpus would have missed the Indonesian rise almost entirely.

Theoretical Implications

Eckert's (2012) three-wave framework still organises variationist sociolinguistics. It anchors Cluster C4 in the term network. The cluster analysis tells a more complicated story. Classroom code-switching research (Cluster C1) and multilingualism research (Cluster C2) operate with a different unit of analysis. Their unit is the speaker's repertoire in an institutional setting, not a discrete variable inside a named language. Their epistemological commitments differ too. They treat the classroom, not the speech community, as the primary site. The two communities share almost no anchor texts: none of the four most-cited documents in the corpus (Table 4) is a variationist work, and C4's share of term occurrences has fallen from 18.9% to 12.2% across the three periods.

Sociolinguistics, then, is a confederation rather than a discipline. A pluralist confederation, with shared institutional venues but diverging theoretical commitments. The translanguaging case sharpens this point. García and Wei (2014) introduced the concept as a coherent alternative to additive bilingualism: discrete bounded languages out, unitary multilingual competence in. The present analysis suggests something has happened to the term since. It bursts in 2020 with a strength of 6.6, but the classroom code-mixing front bursts a year later with a strength of 23.6; in the title vocabulary of the field, translanguaging is now a satellite of a much larger applied programme, not its centre. Peer review has not arrested this drift. Translanguaging should not be abandoned. It needs operationalisation criteria that distinguish theoretical from descriptive use, and it needs them soon.

Methodological Implications

The India Asymmetry has a direct methodological message. Single-database bibliometrics will systematically under-read fields whose output sits in grey literature — dissertations, working papers, national journals not indexed in WoS or Scopus. The implication for future meta-analyses is straightforward. They need multi-source designs. Future reviews of sociolinguistics, and indeed of any field with a substantial non-Anglophone tradition, should at minimum draw on Shodhganga (India), NDLTD (global), and the institutional repositories of major non-Anglophone research universities. The cost is real — assembling and harmonising multiple corpora is more demanding than running a single WoS query. The benefit is also real: the picture changes substantively when the dissertation layer is added.

VOSviewer keyword clustering and CiteSpace burst detection complement each other. Clustering picks up what is established. Bursts pick up what is moving. The 'language and AI' burst illustrates this directly: only three records, far too few to influence cluster assignment, but a measurable spike in time-derivative frequency that any cluster-only analysis would have missed. Combining the two tools should be standard practice in sociolinguistics meta-analysis going forward.

Practical Implications

Three audiences should read these findings differently. For graduate supervisors and programme designers, the regional thematic divergence is a curricular opportunity, not a problem to be solved. The US programmes that centre on variationist methodology would gain from sustained exposure to the language policy frameworks that European programmes prioritise. European programmes would gain from the applied multilingualism toolkit that Indian doctoral training has refined. Indian

programmes would gain from the variationist methods and open-science practices that have travelled less successfully into South Asian institutional networks. Joint doctoral modules and structured exchanges across the three traditions are practical first steps in the direction Adli and Guy (2022) called for.

Journal editors face two specific opportunities. Only ten titles in a corpus of 10,837 name research ethics, and the burst onsets of 'youtube', 'social media' and 'online' (2021–2022) indicate that the data now being collected are increasingly platform-based; submission requirements modelled on those used in communication studies would close this gap with no major infrastructural investment — only an editorial-policy update. The 0.08% share of titles naming open data or replication calls for editorial alignment with practices already standard at Nature, BMJ, and a growing number of social-science journals: data-availability statements at submission, plus either open data or a documented justification for non-disclosure.

Funding bodies face a different problem. The gap matrix in Figure 9 ranks priorities by urgency-impact score, but the structural causes behind each gap differ. Longitudinal research (T3) needs multi-year funding instruments that single-cycle grants cannot supply. Sub-Saharan African sociolinguistics outside South Africa (G1) needs capacity-building infrastructure: doctoral scholarships, open-access publishing support, collaborative grants. The AI gap (T2) is the most time-sensitive of the three. The documentation window for pre-AI language is closing fast. Each of these calls for a different funding instrument, and treating them as interchangeable wastes resources that would otherwise be well targeted.

Limitations

Four limitations bear on the interpretation of the findings. The corpus is restricted to English-language records. This is a methodological necessity for systematic reproducibility, but it excludes a substantial body of sociolinguistics scholarship in Spanish, French, German, Portuguese, Arabic, Hindi, and other languages. Indian work in Hindi, Tamil, and Telugu is entirely absent; Spanish-language sociolinguistics is represented only through its English-language output. A future replication able to incorporate non-English records would substantially refine the geographic picture.

Thematic assignment in the YÖK layer was rule-based, applied to titles only, and performed without a second coder. Inter-rater reliability statistics could therefore not be computed, and 21.1% of theses could not be placed in a named theme. This constraint is deliberate rather than incidental: full-text coding of 185 theses and a multi-coder protocol were outside the resources of a single-author study, and the thematic results in §3.3.3 are presented as descriptive labels for that reason. Two sets of claims are the most sensitive to this constraint, and it is worth naming them rather than leaving the limitation general. The thematic composition of the Turkish layer (Table 8) and the two significant contrasts derived from it (degree and language of writing) depend on the rule set, and a team-based replication using two or more independent coders would allow Krippendorff's α to be computed and would test whether the contrasts are robust to coder identity. The journal-layer clusters are not affected, because they rest on term co-occurrence rather than on coding.

The corpus counts ($n = 10,837$ journal records; $n = 185$ YÖK theses; $n = 547$ open-repository dissertations) reflect the retrieval and screening procedures described in §2 and are subject to revision as OpenAlex is updated. The retrieval tool's method log, the exported record sets, the cleaned corpus and the full test table (21 tests with raw and Benjamini–Hochberg adjusted p -values) provide a public reference point for tracking any deviations from the reported analysis.

Finally, the three data sources vary considerably in temporal depth and coverage. OpenAlex indexes local and non-indexed journals, which inflates the shares of countries with large national journal systems (Indonesia, South Africa, Pakistan) relative to WoS- or Scopus-based studies. YÖK is complete for Turkey but covers no other country. The OpenAlex dissertation layer reaches only 2019–2025 and depends on which repositories expose their holdings; Northern and Western European repositories are over-represented, so the Europe–Asia asymmetry in §3.3.2 measures visibility as much as production. This unevenness should be borne in mind when interpreting the regional comparison, and it is the reason the asymmetry is reported for a matched window only. A parallel qualification applies to the geographic coding itself: 29% of the dissertation layer could not be assigned a country from repository names, and the assigned countries are those of the repository, not necessarily of the degree-awarding institution.

Positioning Summary

According to the table 10 the far-right column collects the contributions claimed in the Introduction and tested across the rest of the manuscript. Three rows do the most analytical work in this table. The 'Design' row marks the methodological inheritance from Marzi et al. (2025) and the divergence from prior sociolinguistics meta-work that has remained purely bibliometric. The 'Global South representation' row crystallises Contribution 4: prior work either treats the Global South implicitly, names its under-representation without measuring it, or critiques it; the present study measures it and finds that for Indonesia the direction has reversed. The 'Source-type asymmetry' row is the cleanest example of what the B-SLR design adds: this is not a finding any prior bibliometric study could have produced, regardless of corpus size or analytical

sophistication, because it requires two source types to be placed on the same regional scale.

Table 10.

Synoptic positioning against the three closest prior studies.

Dimension	Batubara et al. (2024)	Flores Riquelme (2025)	Adli & Guy (2022)	Present Study
Database(s)	Scopus only	WoS + Scopus	Manual journal review	OpenAlex + YÖK + OpenAlex dissertations
Period	2014–2023	2005–2024	1990–2020	2000–2025 / 2019–2025
Unit of analysis	Journal articles	Journal articles	Journal articles	Articles + two dissertation layers
Design	Bibliometric only	Bibliometric only	Narrative manifesto	B-SLR (combined)
Global South representation	Implicit only	Gaps named, not measured	Critique only	Measured (Indonesia, Turkey, Sub-Saharan Africa)
Thematic clusters	4 broad areas	3 clusters	Discussed narratively	9 Leiden clusters + trajectory validation
Gap analysis	Implicit	Explicit, unranked	Manifesto-style	Prevalence-ranked matrix
Source-type asymmetry	Not reported	Not reported	Not reported	Reported (Table 7)
PRISMA compliance	No	No	No	Yes (PRISMA 2020)
Confirms prior	Growth	Translanguaging	Anglophone bias	All three, with qualification
Revises prior	—	—	—	No 2016 acceleration; Anglophone decline; Indonesian rise
New findings	—	—	—	B-SLR; dissertation layers; source-type asymmetry; Turkish profile

Source: Author's own synthesis of Batubara et al. (2024), Flores Riquelme (2025), Adli and Guy (2022) and the present study.

General assessment of the Discussion section

Two judgements are warranted by the analysis. The first concerns the field's external trajectory: sociolinguistics is growing steadily, but the growth is migrating, geographically toward Southeast Asia and thematically toward classroom-based applied work, while the variationist and policy traditions hold their ground without expanding. The second concerns the field's visibility: journal and dissertation records do not show the same geography, and a national system such as Turkey's can be growing rapidly while remaining almost invisible in the journal record. Neither of these judgements supports a story of progressive disciplinary integration. Both support a story in which the discipline's external success masks an internal pluralism that is now deep enough to require explicit acknowledgement. The closing section translates that acknowledgement into a forward agenda.

Conclusion

Eight Empirical Findings

The study established the following findings, each grounded in evidence reported in §3.

- F1. Sociolinguistics output grew at a compound 10.0% per year from 2000 to 2025, from 106 to 1,152 records. There is no 2016 acceleration; the only detectable change of slope is a slight deceleration.
- F2. Nine thematically distinct research communities organise the field. The largest, classroom-oriented code-switching research, rose from 27.2% to 39.7% of term occurrences across three periods, while the variationist community fell from 18.9% to 12.2%.
- F3. Translanguaging is present but not dominant. It bursts in 2020 with a strength of 6.6; the classroom code-mixing front bursts a year later with a strength of 23.6. In the field's title vocabulary, translanguaging is a satellite of a larger applied programme.
- F4. The geography of the field has shifted. The Anglophone core fell from 48.6% to 20.4% of lead-author records between 2000–2009 and 2018–2025; Indonesia rose from 0.5% to 18.2% and is now the second-largest producer. Both trends are significant at $p < .001$.
- F5. Journal and dissertation records do not show the same geography: $\chi^2(4) = 127.5$, $p < .001$, $V = 0.171$ for 2019–2025. Asia produces 43.8% of journal articles but 11.0% of open-repository dissertations; Europe produces 33.7% and 58.6% respectively.
- F6. Three research gaps were ranked as critical on measured prevalence. AI and language change appears in 0.07% of titles, longitudinal designs in 0.15%, and open data or replication in 0.08%, and none of the three shares is rising.
- F7. Turkey shows the source-type asymmetry in miniature. Its thesis output grew 10.2% per year with a structural break in 2013, the share of theses written in Turkish rose from 31% to 64%, yet Turkey supplies 0.8% of lead-author journal records.
- F8. Sign languages occupy 0.40% of the corpus. The share is small but rising: 18 of the 43 records date from 2020 onward, which is why this gap is banded Major rather than Critical in the present analysis.

Six Inferential Claims

Table 11.

Six inferential claims with grounding findings and qualifications.

Claim	Grounding findings	Qualification
C1. Sociolinguistics is growing steadily, not accelerating	F1: compound 10.0%/yr; LRT slope change at 2016 is a deceleration	Applies to the OpenAlex title-search corpus; other search designs may differ
C2. The field's growth is migrating to classroom-based applied work	F2, F3: C1 share 27.2% → 39.7%; code-mixing burst 23.6	Clusters rest on title terms, not author keywords
C3. The Anglophone core is losing its share of the field	F4: 48.6% → 20.4%; Indonesia 0.5% → 18.2%	OpenAlex includes local journals that WoS/Scopus omit; magnitude is source-dependent
C4. Journals and dissertations show different geographies	F5: $\chi^2(4) = 127.5$, $V = 0.171$	Dissertation layer reaches only 2019–2025 and over-represents Europe
C5. AI, longitudinal and open-data work are structurally absent	F6: 0.07%, 0.15%, 0.08% of titles, none rising	Marker terms in titles under-count work that names them only in abstracts
C6. National training systems can grow while remaining invisible in journals	F7: Turkey +10.2%/yr theses vs 0.8% of journal records	Single-country evidence; thematic labels are rule-based

Source: Author's own elaboration based on the findings reported in §3.

According to table 11 the claims are ordered from most secure to most provisional. C1, C3 and C4 rest on quantitative

findings (growth model, share trends, chi-square statistic) that survived Benjamini–Hochberg adjustment across a 21-test family and survive almost any reasonable robustness check. C2 depends on a term-based cluster solution that should be re-examined with author keywords where a database supplies them. C5 and C6 depend on present-state evidence that may shift quickly: the AI front in particular is moving fast enough that the claim's status will need re-evaluation within two years. The table is therefore best read not as a final inventory but as a calibrated set of conclusions that fade in confidence as one moves down the column.

Prioritised Research Agenda

Table 12.

Prioritised nine-item research agenda derived from the gap analysis.

Priority	Action	Horizon
1 — URGENT	Establish baseline corpora documenting pre-AI language use across registers and varieties (Gap T2)	2025–2027
2 — URGENT	Fund longitudinal panel studies of language change with at least ten-year horizons (Gap T3)	2026–2035
3 — HIGH	Mandate data-availability statements and open data or documented non-disclosure at core sociolinguistics journals (Gap M1)	2026–2028
4 — HIGH	Adopt field-wide ethics reporting standards for platform-based data, given the 2021–2022 social-media burst (Gap M2)	2025–2026
5 — HIGH	Support sign-language sociolinguistics so that its rising share (Gap T1) consolidates into a research community	2025–2030
6 — MEDIUM	Build publication infrastructure for national thesis systems that are invisible in the journal record: open-access waivers, editorial mentorship, thesis-to-article pathways (Gap G2, Turkey as the test case)	2025–2030
7 — MEDIUM	Extend capacity-building to Sub-Saharan Africa outside South Africa (Gap G1)	2026–2032
8 — MEDIUM	Integrate computational methods into graduate curricula so that Gap T4's rise is met with trained researchers	2026–2030
9 — ONGOING	Replicate this meta-analysis on a five-year cycle with a dissertation layer that reaches back to 2000	2030, 2035

Source: Author's own elaboration based on the gap matrix (Table 9, Figure 9) and the urgency-impact scoring described in §2.7.

According to table 12 the agenda is intentionally heterogeneous in its time horizons. Priority ordering follows the urgency-impact score derived in §2.7, which combines evidence scarcity (title-marker prevalence), corpus trajectory (whether the share is rising, flat or falling) and structural reversibility — a gap whose conditions are deteriorating receives higher priority than one whose conditions are stable, holding scarcity constant. Priorities 1 and 2 are urgent because their solvability conditions are deteriorating: pre-AI language data becomes unrecoverable as LLM-influenced text becomes more pervasive, and a longitudinal design not started now cannot be started retroactively. Priorities 6 through 8 are medium-horizon items because they require institutional rather than methodological change, and institutional change in graduate education tends to operate on a five- to ten-year cycle. Priority 9 frames the agenda recursively: the field needs to commit to repeated self-assessment if

it is to detect emerging gaps before they become structural deficits.

Closing Statement

Sociolinguistics has long been attentive to whose language counts, who gets to speak, and which linguistic practices the available instruments of analysis make visible or invisible. The present study applies that same attentiveness to the discipline itself. The findings are not technical shortcomings to be acknowledged in passing and forgotten. They are structural patterns of collective attention that the field has the methodological capacity, the theoretical resources, and the professional responsibility to address.

The Bibliometric-Systematic Literature Review design adopted here is itself a methodological argument: that the most productive path to disciplinary self-knowledge runs through systematic evidence synthesis combined with interpretive depth. Neither layer alone tells the full story. The interaction between the two — between the field's countable patterns and its conceptual commitments — is where the most consequential questions live, and where the most actionable recommendations originate.

The map produced by this study is comprehensive enough to be useful, candid about its limitations, and specific enough in its recommendations to be put into practice. The territory it maps — global sociolinguistics scholarship at the midpoint of the 2020s — is more regionally diverse, more theoretically pluralist, and more in need of institutional self-correction than the available meta-literature has previously suggested. Reframing that picture is the principal contribution of this work.

Peer-review: Externally peer-reviewed.

Conflict of Interest: The author has no conflicts of interest to declare.

Financial Disclosure: The author declared that this study has received no financial support.

Use of Artificial Intelligence: The author declares that generative artificial intelligence tools were not used in the study design, data collection, analysis, and interpretation stages.

Hakem Değerlendirmesi: Dış bağımsız.

Çıkar Çatışması: Yazar, çıkar çatışması olmadığını beyan etmiştir.

Finansal Destek: Yazar, bu çalışma için finansal destek almadığını beyan etmiştir.

Yapay Zekâ Kullanımı: Yazar, çalışmanın tasarımı, veri toplama, analiz ve yorumlama aşamalarında üretken yapay zekâ araçlarının kullanılmadığını beyan etmiştir.

References

- Adli, A., & Guy, G. R. (2022). Globalising the study of language variation and change: A manifesto on cross-cultural sociolinguistics. *Language and Linguistics Compass*, 16(5–6), e12452. <https://doi.org/10.1111/lnc3.12452>
- Aria, M., & Cuccurullo, C. (2017). Bibliometrix: An R-tool for comprehensive science mapping analysis. *Journal of Informetrics*, 11(4), 959–975. <https://doi.org/10.1016/j.joi.2017.08.007>
- Batubara, M. H., Nasution, A. K. P., & Nurmalina, N. (2024). Trends in sociolinguistics research in the last decade: Bibliometric analysis. *Edelweiss Applied Science and Technology*, 8(6), 6001–6022. <https://doi.org/10.55214/25768484.v8i6.3314>
- Blommaert, J. (2010). *The sociolinguistics of globalization*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511845307>
- Bucholtz, M., & Hall, K. (2004). Theorizing identity in language and sexuality research. *Language in Society*, 33(4), 469–515. <https://doi.org/10.1017/S0047404504334020>
- Chen, C. (2006). CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. *Journal of the American Society for Information Science and Technology*, 57(3), 359–377. <https://doi.org/10.1002/asi.20317>
- Coulmas, F. (Ed.). (1998). *The handbook of sociolinguistics*. Blackwell.
- Coupland, N. (2007). *Style: Language variation and identity*. Cambridge University Press.
- Crystal, D. (2000). *Language death*. Cambridge University Press.
- Dorian, N. C. (1981). *Language death: The life cycle of a Scottish Gaelic dialect*. University of Pennsylvania Press.
- Eckert, P. (2000). *Linguistic variation as social practice*. Blackwell.
- Eckert, P. (2012). Three waves of variation study: The emergence of meaning in the study of sociolinguistic variation. *Annual Review of Anthropology*, 41, 87–100. <https://doi.org/10.1146/annurev-anthro-092611-145828>
- Fishman, J. A. (1965). Who speaks what language to whom and when? *La Linguistique*, 1(2), 67–88.
- Fishman, J. A. (1991). *Reversing language shift*. Multilingual Matters.
- Flores Riquelme, A. C. (2025). Mapping the field: A bibliometric literature review of language attitudes in sociolinguistic studies. *Transformatika*, 9(1), 1–28.
- García, O., & Wei, L. (2014). *Translanguaging: Language, bilingualism and education*. Palgrave Macmillan.

- Grenoble, L. A., & Whaley, L. J. (2006). *Saving languages: An introduction to language revitalization*. Cambridge University Press.
- Gumperz, J. J., & Hymes, D. (Eds.). (1972). *Directions in sociolinguistics*. Holt, Rinehart and Winston.
- Hornberger, N. H. (Ed.). (2003). *Continua of biliteracy*. Multilingual Matters.
- Inflibnet Centre. (2024). *Shodhganga: A reservoir of Indian theses*. <https://shodhganga.inflibnet.ac.in>
- Labov, W. (1966). *The social stratification of English in New York City*. Center for Applied Linguistics.
- Labov, W. (1972). *Sociolinguistic patterns*. University of Pennsylvania Press.
- Marzi, G., Balzano, M., Caputo, A., & Pellegrini, M. M. (2025). Guidelines for bibliometric-systematic literature reviews. *International Journal of Management Reviews*, 27(1), 81–103. <https://doi.org/10.1111/ijmr.12381>
- Milroy, L., & Gordon, M. (2003). *Sociolinguistics: Method and interpretation*. Blackwell.
- Mongeon, P., & Paul-Hus, A. (2016). The journal coverage of Web of Science and Scopus: A comparative analysis. *Scientometrics*, 106(1), 213–228. <https://doi.org/10.1007/s11192-015-1765-5>
- Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., ... Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, 372, n71. <https://doi.org/10.1136/bmj.n71>
- Satyanath, S. (2021). Genealogies of sociolinguistics in India. *Journal of Sociolinguistics*, 25(5), 762–784. <https://doi.org/10.1111/josl.12496>
- van Eck, N. J., & Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics*, 84*(2), 523–538.

Supplementary Material

Supplementary Table S1.

Three-zone Bradford composition of the journal corpus ($n = 10,837$; 4,456 venues). Zone boundaries follow the equal-thirds convention (§2.5); the 40 most productive journals are listed with their zone, and the composition of Zones 2 and 3 is summarised in the final rows.

Bradford zone	Rank	Journal	Records (n)	Cumulative % of corpus
Zone 1 (core)	1	Journal of Multilingual and Multicultural Development	227	2.1%
	2	Language Policy	224	4.2%
	3	Current Issues in Language Planning	140	5.5%
	4	International Journal of the Sociology of Language	137	6.7%
	5	International Journal of Multilingualism	129	7.9%
	6	Journal of Sociolinguistics	110	8.9%
	7	International Journal of Bilingualism	103	9.9%
	8	Language in Society	102	10.8%
	9	International Journal of Bilingual Education and Bilingualism	77	11.5%
	10	Sociolinguistic Studies	71	12.2%
	11	Multilingua	65	12.8%
	12	Languages	65	13.4%
	13	Language Problems & Language Planning	62	14.0%
	14	Sociolinguistica - International Yearbook of European Sociolinguistics / Internationales Jahrbuch für europäische Soziolinguistik	47	14.4%
	15	Bilingualism Language and Cognition	41	14.8%
	16	European Journal of Language Policy	38	15.1%
	17	Language and Education	37	15.5%
	18	Journal of Language Identity & Education	35	15.8%
	19	Language Matters	34	16.1%
	20	Sustainable Multilingualism	34	16.4%
	21	Theory and Practice in Language Studies	33	16.7%
	22	Southern African Linguistics and Applied Language Studies	32	17.0%
	23	Työväentutkimus Vuosikirja	32	17.3%
	24	World Englishes	31	17.6%
	25	Language & Communication	30	17.9%
	26	Applied Linguistics	30	18.1%
	27	Frontiers in Psychology	29	18.4%
	28	Language and Linguistics	28	18.7%

		Compass		
	29	Journal of Historical Sociolinguistics	27	18.9%
	30	South African Journal of African Languages	26	19.2%
	31	Langage et société	26	19.4%
	32	International Journal of Applied Linguistics	26	19.6%
	33	Modern Linguistics	26	19.9%
	34	Language	24	20.1%
	35	Studies in Second Language Acquisition	23	20.3%
	36	Applied Linguistics Review	23	20.5%
	37	Critical multilingualism studies	23	20.7%
	38	American Speech	22	20.9%
	39	International Multilingual Research Journal	22	21.1%
	40	ELYSIAN JOURNAL English Literature Linguistics and Translation Studies	22	21.3%
Zone 1 (rest)	41–138	98 further journals contributing 12 to 43 records each	1,528	33.3%
Zone 2	139–1,227	1,089 journals contributing 2 to 11 records each	3,616	66.7%
Zone 3	1,228–4,456	3,229 journals contributing a single record each	3,613	100.0%

Note. Venue names are the OpenAlex source names at the time of retrieval (27 September 2026). Zone 1 comprises 138 journals and 3,608 records (33.3% of the corpus); Zone 2 comprises 1,089 journals and 3,616 records (33.4%); Zone 3 comprises 3,229 journals and 3,613 records (33.3%). The Zone 3 : Zone 1 journal ratio ($3,229 : 138 = 23.4$) indicates the very high scatter characteristic of a multidisciplinary field indexed beyond the WoS–Scopus perimeter.

Source: Author's own elaboration; Bradford zone analysis of the OpenAlex journal corpus following Aria and Cuccurullo (2017).