WordNet und der Atlas der Sprache

Schlussbericht des Forschungsstrangs und kanonischer Import

Deutsch und English · Version 2.0 · 11. September 2026

Autor: Michael Reich

Kanonische Bezugsbasis: ATLAS 250 · Sites 213 · Datenmodell 1.6.0

Forschungsfreeze: d3c9d1c165ee15896e029cb0a8f4602ae96016f1

Gate SHA-256: 6f4df73fc3b19c984f1b74030259d7a23c2a5b0bd0f8505bab64946f0f66f4ea

Kanonischer Paket-Hash: 283b79b8e42335640d82b9d6b39fc27f27431cb2066ff0666af9fabfca7872d5

Empfohlene Zitierweise Reich, Michael (2026): WordNet und der Atlas der Sprache. Schlussbericht des Forschungsstrangs und kanonischer Import. Version 2.0. Atlas der Sprache. https://doi.org/10.5281/zenodo.22713535

Status Forschung, redaktionelles Gate, Rebase, kanonischer Import und lokaler Readback abgeschlossen.

Teil I Deutsch

Abstract

Der Bericht dokumentiert den vollständigen Abgleich der ursprünglich 1.028 kanonischen Begriffe des Atlas der Sprache mit Princeton WordNet 3.0 und seine kontrollierte Übernahme in ATLAS 250. Das Erstscreening ergab 196 Single-Sense-, 169 polyseme, 38 hochambige und 625 direkte NO_MATCH-Fälle. Review 11 schloss die letzten zehn Dossierfälle. Im manuellen Audit der 196 scheinbar eindeutigen Treffer wurden 155 Crosswalks bestätigt und 41 Falschzuordnungen verworfen. Für die 625 NO_MATCH-Fälle kam ein gezieltes Hochrisiko-Screening statt einer schematischen Vollprüfung zum Einsatz. Das redaktionelle Hauptatlas-Gate umfasste 284 eindeutige Konzepte und 320 Provenienzoperationen; daraus wurden 307 neue Sense-Key-Crosswalks importiert. Nach vier kanonischen Dublettenzusammenführungen enthält ATLAS 250 insgesamt 342 WordNet-Crosswalks zu 301 kanonischen Konzepten. Vier Legacy-IDs bleiben über dauerhafte Redirects auflösbar. Von 17 gesonderten Lexikalisierungs- und Scope-Befunden wurden 14 Labelentscheidungen umgesetzt und drei Scope-Fragen vertagt; der früher bestätigte Fall Rückbildung wurde zusätzlich von Reduction auf Back-formation korrigiert. WordNet bleibt ein externer Sense- und Interoperabilitätsanker. Es ersetzt weder die Atlasdefinitionen noch die interne Ontologie.

Schlüsselwörter: Princeton WordNet 3.0; Sense-Key; lexikalische Ontologie; Crosswalk; Falschfreunde; Coverage Gap; Forschungsdatenprovenienz; Atlas der Sprache; kanonischer Import.

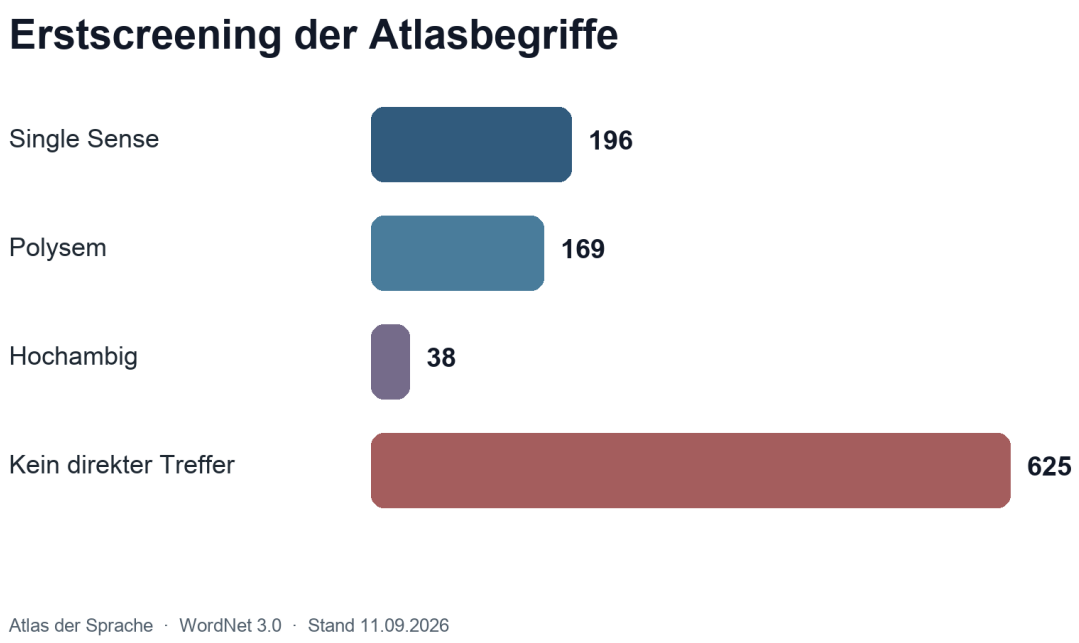

Abbildung 1 Erstscreening aller 1.028 Atlasbegriffe

1 Forschungsfrage und Ergebnisgrenze

Der Forschungsstrang beantwortet drei voneinander zu trennende Fragen:

1. Ist ein englisches Atlas-Lemma in WordNet 3.0 überhaupt als Suchform vorhanden?
2. Wenn es vorhanden ist: Welcher konkrete WordNet-Sense bewahrt die kanonische Atlasdefinition?
3. Wenn kein direkter Treffer vorliegt: Fehlt WordNet tatsächlich der Fachbegriff, oder verdeckt eine problematische englische Atlas-Lexikalisierung einen vorhandenen Sense?

Die Untersuchung prüft damit **lexikalisch-semantische Interoperabilität**, nicht ontologische Identität. WordNet organisiert englische Substantive, Verben, Adjektive und Adverbien in Synsets und verknüpft diese durch lexikalische und konzeptuelle Relationen; Synsets tragen gemeinsame Glosses und gegebenenfalls Beispiele (Princeton University, About WordNet). Der Atlas enthält demgegenüber zahlreiche historische, rhetorische, linguistische, medientheoretische und kulturspezifische Concepts, deren Definitionen absichtlich breiter oder anders granularisiert sein können.

Ein positiver Crosswalk bedeutet daher: *Dieser WordNet-Sense ist ein belastbarer externer Anker für den semantischen Kern des Atlasconcepts.* Er bedeutet nicht: *Beide Datenmodelle oder Definitionen sind vollständig deckungsgleich.* Ebenso ist `NO_MATCH` kein Qualitätsurteil über den Atlas. Es kann eine echte fachliche Abdeckungslücke von WordNet markieren.

2 Daten und Identifikatoren

2 1 Atlas Grundgesamtheit

Die Grundgesamtheit bilden 1.028 Concepts aus dem kanonischen `concepts.jsonl` der oben ausgewiesenen Main-Basis. Für jedes Concept wurden ID, deutscher Term, englische Übersetzung, Kategorie und beide Definitionen bezogen. Die ID, nicht die Wortform, ist die Identitätsachse. Das verhindert, dass Homonyme oder doppelt überlieferte Fachfiguren allein wegen gleicher Zeichenketten zusammenfallen.

2 2 Externe Ressource

Verwendet wurde Princeton WordNet **3.0**. Die Version wird ausdrücklich festgehalten, weil Synset-Offsets versionsabhängig sind. Für dauerhafte Verweise speichert der Atlas primär Sense-Keys; das WordNet-Referenzhandbuch bezeichnet den Sense-Key als bevorzugte Repräsentation eines konkreten Sense und erläutert, dass Sense-Nummern und Synset-Offsets zwischen Versionen variieren können (Princeton WordNet, `senseidx(5WN)`). Zusätzlich wurden POS, achtstelliger Synset-Offset, Synsetname und Gloss dokumentiert, um jeden Review nachvollziehbar zu machen.

WordNet 3.0 umfasst laut offizieller Statistik 117.659 Synsets und 206.941 Wort-Sense-Paare in den vier offenen Wortklassen (Princeton WordNet, `wnstats(7WN)`). Diese Größe darf nicht als Garantie fachsprachlicher Vollständigkeit missverstanden werden: Die Ressource modelliert lexikalisierte englische Senses, nicht jede fachliche Ontologie oder jedes historische Terminologiesystem.

3 Methodik

3 1 Vollständiges Erstscreening

Alle 1.028 Concepts wurden batchweise gegen normalisierte englische Lemmaformen in WordNet 3.0 geprüft. Mehrwortausdrücke wurden in der WordNet-Konvention mit Unterstrichen getestet; Groß-/Kleinschreibung und reguläre morphologische Varianten wurden normalisiert. Die Screeningklassen waren:

Klasse	n	Anteil an 1.028	Bedeutung
`SINGLE_SENSE`	196	19,1 %	genau ein lemma-genauer WordNet-Sense
`POLYSEMOUS`	169	16,4 %	mehrere plausible Kandidaten in begrenzter Zahl

Klasse	n	Anteil an 1.028	Bedeutung
HIGH_AMBIGUITY	38	3,7 %	hohe Lemma-Mehrdeutigkeit; stärkeres Ausschlussreview nötig
NO_MATCH	625	60,8 %	kein direkter normalisierter Lemmatreffer

Das Screening ist vollständig, aber noch keine Bedeutungsentscheidung. WordNet selbst trennt Zeichenketten von konkreten Senses; gerade deshalb wäre die bloße Lemmaexistenz als Importkriterium methodisch unzulässig (Princeton University, About WordNet).

3 2 Manuelles Sense Review

Für jeden reviewten Fall wurde die kanonische Atlasdefinition gegen **jede** lemma-genaue WordNet-Gloss verglichen. Geprüft wurden:

- semantischer Kern: Bezeichnet der Sense denselben Gegenstands- oder Relationskern?
- Domain: Gehört der Sense zur richtigen Fachdomäne, etwa Rhetorik statt Biologie oder Narratologie statt Musik?
- Granularität: Ist der WordNet-Sense breiter, enger oder nur ein Nachbarbegriff?
- ontologischer Typ: Prozess, Ergebnis, Relation, Eigenschaft, Person, Zeichen oder Praxis;
- POS und Lexikalisierungsform: nominales Concept gegenüber Verb, Adjektiv oder derivationalem Nachbarn;
- Mehrfachbedarf: Deckt ein einzelner Synset die Atlasdefinition oder sind mehrere Senses erforderlich?

Die Entscheidungsklassen lauten:

- EXACT_OR_NEAR_MATCH: der semantische Kern bleibt erhalten; dokumentierte Reichweitendifferenzen sind zulässig;
- MULTIPLE_SYNSETS_REQUIRED: die Atlasdefinition bündelt zwei oder mehr von WordNet getrennte Senses;
- NO_ADEQUATE_MATCH: kein vorhandener Kandidat trägt die Definition ausreichend.

Negativfälle wurden zusätzlich als allgemeine Abdeckungslücke, WORDNET_FALSE_FRIEND, EXTERNAL_HOMOGRAPH_NO_DOMAIN_SENSE oder spezifischer Lexikalisierungs-/Scope-Befund klassifiziert. Die Gloss wurde als Prüfevidenz, nicht als bloßes Stringmaterial verwendet. WordNet definiert eine Gloss als synsetbezogene Definition und Beispiele; der Sense ist jeweils einem eigenen Synset zugeordnet (Princeton WordNet, Glossary).

3 3 Review 11 Abschluss des Restdossiers

Review 11 enthielt genau die zehn nach Review 10 noch offenen, nicht bereits kanonisch gebundenen Dossierfälle. Alle Kandidatensenses wurden einzeln ausgeschlossen oder ausgewählt:

Atlasconcept	Entscheidung	ausgewählter WordNet-3.0-Sense
Tautologie	exact/near	tautology%1:10:00::
Zirkumlocution	exact/near	circumlocution%1:10:01::
Meiosis	exact/near	meiosis%1:10:00::
Sentenz	exact/near	maxim%1:10:00::
Kryptographie	exact/near	cryptography%1:04:00::

Atlasconcept	Entscheidung	ausgewählter WordNet-3.0-Sense
Layout	exact/near	`layout%1:09:00::`
Stenografie	exact/near	`shorthand%1:10:00::`
Typografie	exact/near	`typography%1:04:00::`
Semantik	exact/near	`semantics%1:09:00::`
Mimik	multi-synset	`facial_expression%1:10:00::` + `facial_expression%1:07:00::`

Der Mehrfachfall *Mimik* ist methodisch aufschlussreich: Die Atlasdefinition verbindet ausgeführte Gesichtsbewegung und ausgedrücktes Gefühl beziehungsweise Haltung. WordNet trennt Ausdrucksbewegung und Ausdrucksgehalt. Die Mehrfachbindung bewahrt diese Differenz, statt sie durch eine scheinbar einfache Eins-zu-eins-Zuordnung zu verlieren.

3 4 Manueller Single Sense Audit

Die 196 scheinbar eindeutigen Fälle wurden nicht automatisch akzeptiert. Für jeden Datensatz wurden Atlasdefinition, einziger WordNet-Sense, Gloss, POS, Synset-Offset und Sense-Key in einem festen Dossier verglichen. Ergebnis:

Ergebnis	n	Anteil an 196
`EXACT_OR_NEAR_MATCH`	155	79,1 %
`NO_ADEQUATE_MATCH`	41	20,9 %

Die Zurückweisungsquote von 20,9 % zeigt, dass **Monosemie innerhalb WordNet keine semantische Eindeutigkeit gegenüber einer externen Fachontologie garantiert**. Unter den 41 Negativfällen waren 21 klare WordNet-Falschfreunde, zehn Homographen ohne passenden Domain-Sense und zehn sonstige Reichweiten- oder Ontotypabweichungen.

Beispiele sind *Distinctive feature* (alltagssprachlich auffälliges Merkmal statt phonologisches distinktives Merkmal), *Umlaut* (Diakritikum statt Lautwandel), *Polyphony* (musikalische Stimmen statt narrativer Mehrstimmigkeit), *Metalepsis* (metonymische Kette statt Ebenenüberschreitung), *Applicative* (alltagssprachlich anwendbar statt grammatische Applikativkonstruktion) und *Postposition* (allgemeines Nachstellen beziehungsweise Adposition statt Rechtsversetzung). Diese Fälle wären bei einer automatischen Single-Sense-Annahme als systematische Fehlanker in den Atlas gelangt.

3 5 Gezieltes `NO_MATCH` Lexikalisierungs QA

Die 625 `NO_MATCH`-Fälle wurden bewusst **nicht** vollständig einzeln reviewt. Stattdessen wurde ein zweistufiges, recall-orientiertes Risiko-Screening eingesetzt:

1. Generierung von Oberflächenvarianten, Wortbildungs- und Schreibnachbarn sowie definitionell ähnlichen WordNet-Glosses;
2. manuelle Prüfung aller 30 High-Information-Fälle mit enger Schreibnähe und/oder starkem Definitionssignal sowie vier zusätzlicher Compound-Proben.

Die maschinelle Kandidatenliste umfasste 295 Fälle. Diese Zahl ist kein Korrekturbefund: 277 Hinweise entstanden bereits durch irgendeine WordNet-Oberflächenvariante eines Teilworts. Generische Kopfworttreffer wie *language*, *formation*, *world* oder *voice* erzeugen hohe Recall-Werte, bewahren den fachlichen Compound aber meist nicht. Manuell entschieden wurden deshalb 34 gezielt ausgewählte Fälle.

Die 34 Entscheidungen verteilten sich auf 13 Falschfreunde, fünf bloße Kopfworttreffer, vier echte Compound-Coverage-Gaps, drei derivationale Nachbarn, drei übersehene exakte Kopfwort-Senses, zwei orthographische Variantenmatches sowie einzelne POS-, Scope- und Normalisierungsfälle. Acht Fälle ergaben belastbare neue Crosswalk-Vorschläge:

- *Antonomasia* → WordNets Schreibvariante `antinomasia%1:10:00::` (zwei getrennte Atlas-IDs mit eigenständiger Provenienz);
- *Grammatical number* → `number%1:10:05::`;
- *Lexical tone* → `tone%1:07:02::`;
- *Numeral classifier* → `classifier%1:10:00::`;
- *Speaking in tongues* → `speak_in_tongues%2:32:00::` als POS-/Nominalisierungsvariante;
- *Lexicalization* → `lexicalization%1:22:00::`;
- *Word coinage* → zwei Senses für Ergebnis und Handlung.

Der bekannte Fehlerfall *Rückbildung* → *Reduction* lag nicht in der 625er-Population, weil er bereits in einem früheren Polysemie-Review diagnostiziert worden war. Er wurde im Schlussaudit als Lexikalisierungsbefund erneut in das Hauptatlas-Gate aufgenommen: fachlich angemessen ist *Back-formation*. Neu hinzu kommt *Lexikalisierung* → *Lexicization*: Die englische Normalform *Lexicalization* ist sowohl fachsprachlich als auch in WordNet belegt. Beide Änderungen wurden als getrennte Atlasentscheidungen behandelt. ATLAS 250 führt Rückbildung nun als Back-formation und Lexikalisierung als Lexicalization; die früheren Formen bleiben als Varianten und damit als nachvollziehbare Provenienz erhalten.

4 Gesamtergebnisse

4 1 Eingefrorener Forschungsledger

Manuelle Klasse	n	Anteil an 1.028	Anteil an 436 Reviewten
`EXACT_OR_NEAR_MATCH`	272	26,5 %	62,4 %
`MULTIPLE_SYNSETS_REQUIRED`	26	2,5 %	6,0 %
`NO_ADEQUATE_MATCH`	138	13,4 %	31,7 %
nicht manuell reviewt	592	57,6 %	—
gesamt	**1.028**	**100 %**	—

Damit besitzen 298 manuell reviewte Konzepte einen tragfähigen Einzel- oder Mehrfachanker. Nach Rebase, Redaktion und kontrolliertem Import enthält ATLAS 250 insgesamt 342 WordNet-Sense-Key-Crosswalks zu 301 kanonischen Konzepten. Die Differenz zu den 298 positiven Ledgerentscheidungen erklärt sich durch vorbestehende Paketprovenienz und bereits kanonisch gebundene Fälle. Forschungsklasse, Paketprovenienz und kanonischer Importstatus bleiben in den Dateien getrennt.

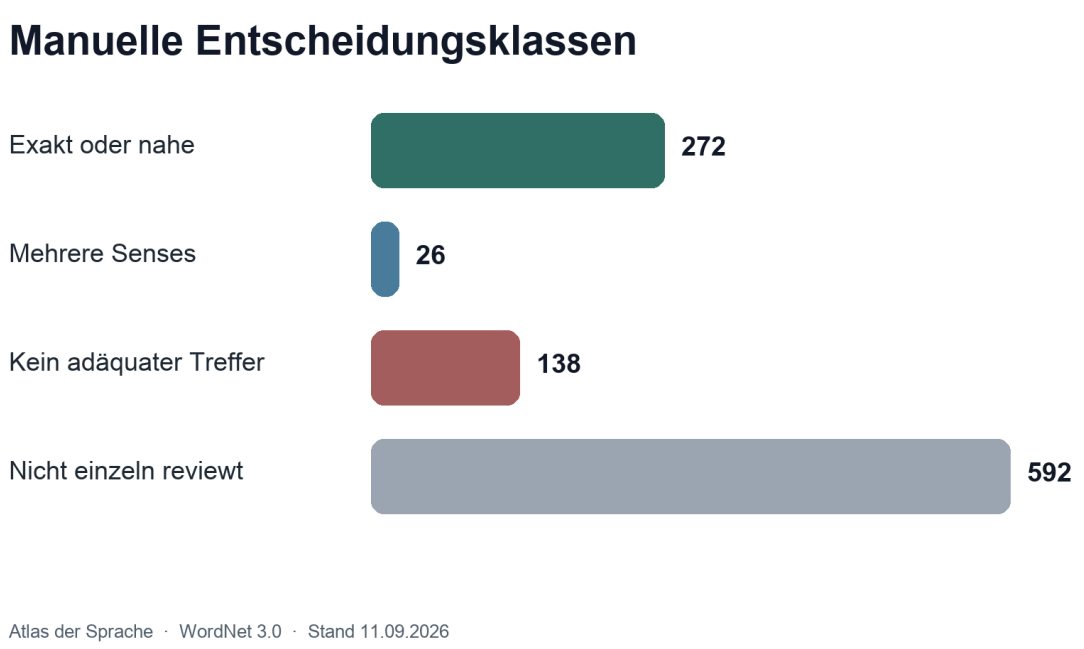

Abbildung 2 Eingefrorene manuelle Entscheidungsklassen

4 2 Verwertbare wissenschaftliche Befunde

Erstens: Lemma-Treffer und Sense-Treffer sind empirisch verschiedene Größen. Schon bei genau einem WordNet-Sense mussten 41 von 196 Fällen zurückgewiesen werden.

Zweitens: Die größten Fehlerrisiken liegen nicht bei zufälligen Tippfehlern, sondern bei systematischen Domainverschiebungen: Musik ↔ Narratologie, Alltagsbedeutung ↔ Grammatik, Medizin/Biologie ↔ Rhetorik und materielles Zeichen ↔ sprachhistorischer Prozess.

Drittens: Mehrworttermini erzeugen eine charakteristische Unterabdeckung. WordNet kennt häufig das Kopfwort (*number*, *tone*, *classifier*, *head*), aber nicht den Atlas-Compound. Ein Kopfwort kann exakter Sense-Anker, bloßer Hyperonym oder Falschfreund sein; nur die Glossprüfung trennt diese Fälle.

Viertens: Der Atlas ist in historischen rhetorischen Figuren und neueren linguistischen Spezialbegriffen systematisch granularer. Begriffe wie *Distinctio*, *Conglobatio*, *Significatio*, narratologische *Metalepsis* oder *discourse world* markieren keine isolierten Datenfehler, sondern modellbedingte Coverage Gaps.

Fünftens: Mehrfachbindungen sind keine Notlösung, sondern ein positives Modellierungsergebnis. Sie zeigen, wo ein Atlasconcept mehrere von WordNet getrennte lexikalisierte Senses integriert, etwa Ausdrucksbewegung und Ausdrucksgehalt bei *Mimik* oder Ergebnis und Handlung bei *Word coinage*.

5 Coverage Gaps

Die 625 direkten `NO_MATCH`-Fälle bestehen aus 326 spezialisierten Einwortformen, 270 Mehrwortausdrücken und 29 Bindestrichformen. Daraus folgen drei Haupttypen:

1. **Fachhistorische Lücke:** lateinische oder griechische Rhetoriktermini, deren alltagssprachliche Ableitungen in WordNet andere Senses tragen.
2. **Kompositionelle Lücke:** der englische Compound ist terminologisch etabliert, WordNet führt aber nur Kopf oder Bestandteile.
3. **Modellgranularitätslücke:** WordNet repräsentiert einen allgemeineren lexikalisierten Sense, während der Atlas eine sprachwissenschaftliche Relation, Praxis oder historische Figur trennt.

Diese Lücken sind verwertbar: Sie liefern eine empirisch abgegrenzte Kandidatenmenge für spätere Fachlexikon-Crosswalks, eine mögliche WordNet-Erweiterungsstudie oder Atlas-interne Domainqualifikationen. Sie rechtfertigen jedoch keine Ersetzung eines präzisen Atlasterms durch ein allgemeineres WordNet-Lemma.

6 Atlas Lexikalisierungs und Scope Befunde

Der Schlussaudit übergibt 17 getrennte redaktionelle Befunde. Besonders handlungsnah sind:

- *Rückbildung*: *Back-formation* statt *Reduction*;
- *Lexikalisierung*: *Lexicalization* statt *Lexicization*;
- *Lokution*: Prüfung von *locutionary act*;
- *Reparatur*: *conversational repair* statt *remediation*;
- *Kongruenz*: *grammatical agreement*;
- *Rechtsversetzung*: *right dislocation* statt *postposition*;
- *Applikativ*: Domainqualifikation *applicative construction*;
- *Dekorum*, *Synaesthesia*, *Parabola* und *Polyphony*: fachliche Qualifikatoren prüfen;
- *Prolepsis* und *Hyperbaton*: mögliche Scope-Aufspaltung beziehungsweise Subtypisierung prüfen.

Die 17 Fälle wurden einzeln redaktionell entschieden. Vierzehn englische Labels wurden präzisiert oder normalisiert; die historisch oder fachlich notwendigen Ausgangsformen bleiben als Varianten erhalten. Gemeinsinn, Prolepsis und Hyperbaton bleiben vertagt, weil eine bloße Umbenennung ihre offene Begriffsgeschichte oder Scope-Frage verdecken würde. Der bereits in Review 08 bestätigte Fall Rückbildung wurde zusätzlich als Back-formation umgesetzt. Damit enthält ATLAS 250 fünfzehn angenommene Labelkorrekturen und drei ausdrücklich sichtbare Vertagungen.

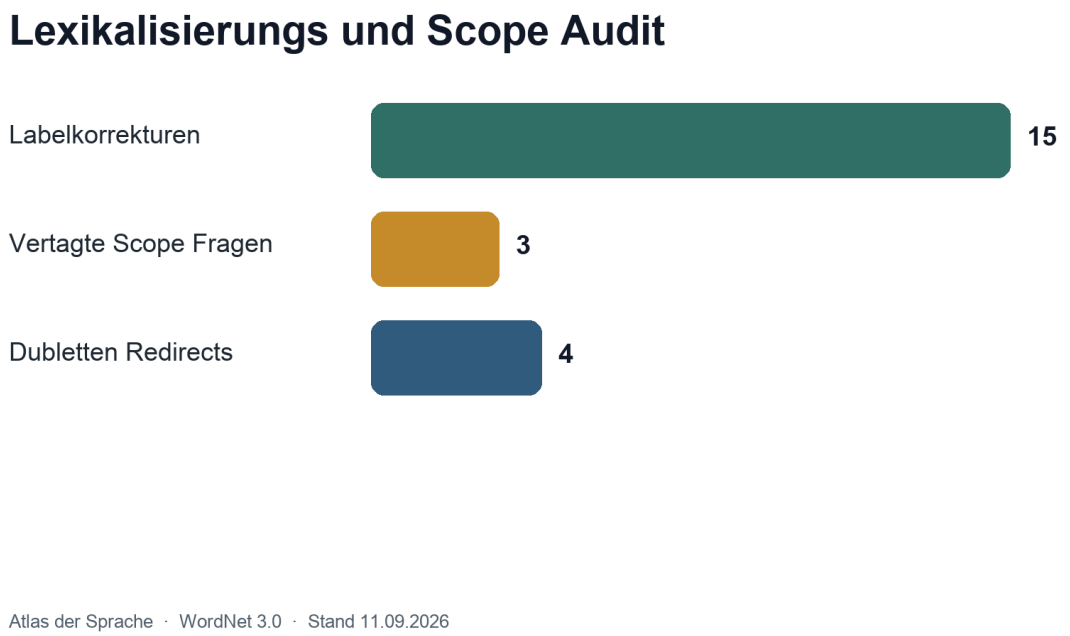

Abbildung 3 Redaktionelle Lexikalisierungsentscheidungen

7 Kanonisches Hauptatlas Gate und Import

Abbildung 4 Konsolidierung und kanonischer Import des Gate 284

Die 137 vorbestehenden Entscheidungen blieben beim Rebase als eigene Eingangskohorte erkennbar. Zusammen mit 173 Schlussauditoperationen und 26 überlappenden Konzepten ergab sich ein Gate von 284 eindeutigen Konzepten beziehungsweise 320 vollständig erhaltenen Provenienzoperationen. Jede Konzeptentscheidung wurde im Ledger `WORDNET_CANONICAL_GATE_284_DECISIONS.json` mit Quellpaket, Operation, Sense-Key und Ziel-ID dokumentiert.

Das angenommene Paket `ATLAS-INBOX-WORDNET-GATE-284-ACCEPTED-V2` atomisiert Mehrfachbindungen und dedupliziert identische Provenienz. Dadurch entstanden 307 neue typisierte Sense-Key-Crosswalks. Die Freigabe stützte sich nicht auf Lemmagleichheit oder Monosemie, sondern ausschließlich auf die vorangegangenen Gloss-, Domain-, Granularitäts- und Ontotypprüfungen.

Vier im Gate bestätigte Benennungsdubletten wurden getrennt vom additiven Crosswalk-Import transaktional zusammengeführt:

- `AT-172-0001` Anadiplosis wurde `AT-RHE-W0003` Anadiplose zugeführt.
- `AT-HAR5-0001` Epanorthosis wurde `AT-119-002` Epanorthose zugeführt.
- `AT-RES6-0011` Symploke wurde `AT-HAR7-0008` Symploce zugeführt.
- `AT-CORE-239-042` Portmanteau wurde `AT-METZ3-0006` Mot-valise zugeführt.

Die Migration übertrug Profile, Sprachformen, Beispiele, Evidenzen und externe Verknüpfungen auf die Survivor-IDs, deduplizierte sechs bereits vorhandene Relationsidentitäten und zwei bereits vorhandene FrameNet-Identitäten und erzeugte vier dauerhafte Redirects. Keine Legacy-ID wurde neu vergeben. Der kanonische Endstand umfasst 1.024 Konzepte, 3.427 gerichtete Relationen, 1.739 Evidenzdatensätze, 872 typisierte externe Crosswalks, neun typologische Crosswalks und vier Redirects. Die Evidenzsumme schließt die bereits im reconcilierten Basisstand Atlas 8.15.6 enthaltenen 21 HTE-Verknüpfungen ein und ist daher nicht WordNet zuzuschreiben.

Der Import wurde auf einem frischen Main-Stand aufgebaut. Der vollständige Migrationslauf endete mit `PRAGMA integrity_check = ok`, ohne Fremdschlüsselfehler und ohne offene Relationen. Der verifizierte Checkpoint `atlas-250-wordnet-gate-284-checkpoint` bindet den Paket-Hash `283b79b8e42335640d82b9d6b39fc27f27431cb2066ff0666af9fabfca7872d5`. Der kanonische Export reproduziert 342 WordNet-Verknüpfungen zu 301 Konzepten.

8 Reproduzierbarkeit und Qualitätssicherung

Der Daten-Freeze umfasst:

- vollständige Screeningbatches für 1.028 Concepts;
- Eingabedossiers und Entscheidungen für Reviews 1–11;
- den 196-zeiligen Single-Sense-Datensatz mit Sense-Key, POS, Offset und Gloss;
- die 625er-`NO_MATCH`-Typologie, die 295er Recall-Liste und 34 manuelle QA-Entscheidungen;
- den abgeleiteten zeilengleichen Forschungsledger (`CSV`, `JSONL`, Summary);
- schema-validierte `needs-review`-Pakete;
- das kanonische Gate als `JSON`, `CSV` und lesbare Dokumentation.

Die Zählungen sind durch Assertions geschützt: 1.028 eindeutige Concept-IDs, 1.028 Screeningzeilen, 196 Single-Sense-Entscheidungen, zehn Review-11-Fälle, 625 `NO_MATCH`-Fälle und die Gate-Mengenlogik 137 + 173 − 26 = 284. Die offizielle WordNet-Lizenz gestattet Nutzung und Weitergabe unter den dort genannten Bedingungen; Provenienz und Copyright-Hinweis sind beizubehalten (Princeton WordNet, License and Commercial Use).

9 Limitationen

- Es wurde WordNet 3.0 untersucht; neuere oder alternative Wordnets können andere Abdeckung bieten.
- Definition-gegen-Gloss-Review ist kuratiert und nachvollziehbar, aber kein Inter-Annotator-Experiment. Cohen-Kappa oder vergleichbare Übereinstimmungsmaße liegen nicht vor.
- Nur 34 der 625 `NO_MATCH`-Fälle erhielten eine manuelle Einzelentscheidung. Die übrigen 591 sind explizit unreviewte Coverage-Population, nicht bestätigte Negativfälle.
- `EXACT_OR_NEAR_MATCH` umfasst dokumentierte Reichweitendifferenzen. Für Anwendungen, die strikte logische Äquivalenz verlangen, ist eine engere Unterklasse erforderlich.
- Die drei vertagten Scope-Fälle Gemeinsinn, Prolepsis und Hyperbaton bleiben außerhalb dieses Releases offen und dürfen nicht als stillschweigend gelöst gelten.

10 Schlussfolgerung

Der WordNet-Forschungsstrang und das getrennte kanonische Import-Gate sind abgeschlossen. Das vollständige Screening, Review 11, der manuelle Single-Sense-Audit, das gezielte NO_MATCH-Lexikalisierungs-QA, der zeilenweise Forschungsfreeze und die redaktionelle Entscheidung aller 284 Gate-Konzepte sind reproduzierbar dokumentiert. ATLAS 250 übernimmt 307 neue Sense-Key-Verknüpfungen, konsolidiert vier bestätigte Benennungsdubletten und bewahrt ihre Legacy-IDs als Redirects.

Die wichtigste methodische Folgerung bleibt bestehen: Die Anzahl der WordNet-Kandidaten ist kein Ersatz für semantische Prüfung. Selbst bei genau einem Sense waren 20,9 Prozent der Kandidaten für die externe Verknüpfung ungeeignet. Für die 625 direkten NO_MATCH-Fälle war ein gezieltes Risiko-Screening ergiebiger als eine schematische Vollannahme oder Vollverwerfung. WordNet erhöht damit die Interoperabilität des Atlas, ohne dessen fachliche Granularität, historische Terminologie oder interne Relationslogik zu bestimmen.

Der öffentliche Release 9.0.0 friert diesen Zustand ein. Die nächste Atlasversion bleibt durch die gemeinsame Strangübergabe gesperrt, bis CLICS, FrameNet, WALS, chinesische Rhetorik und die öffentliche Site den neuen kanonischen Snapshot ausdrücklich abgeglichen haben.

Part II English

Abstract

This report documents the complete comparison of the original 1,028 canonical Atlas of Language concepts with Princeton WordNet 3.0 and its controlled integration into ATLAS 250. Initial screening produced 196 single-sense, 169 polysemous, 38 high-ambiguity and 625 direct NO_MATCH cases. Review 11 closed the final ten dossier cases. Manual auditing of the 196 apparently unambiguous hits accepted 155 crosswalks and rejected 41 unsafe mappings. The 625 NO_MATCH cases were handled through targeted high-risk review rather than exhaustive item-by-item adjudication. The editorial main-atlas gate comprised 284 unique concepts and 320 provenance operations, yielding 307 new sense-key crosswalks. After four canonical duplicate merges, ATLAS 250 contains 342 WordNet crosswalks for 301 canonical concepts and four persistent legacy-ID redirects. Fourteen of 17 separate lexicalization findings were applied while three scope questions were deferred; the previously established Rückbildung case was additionally corrected from Reduction to Back-formation. WordNet remains an external sense and interoperability anchor. It does not replace Atlas definitions or the internal ontology.

Keywords: Princeton WordNet 3.0; sense key; lexical ontology; crosswalk; false friends; coverage gap; research provenance; Atlas of Language; canonical import.

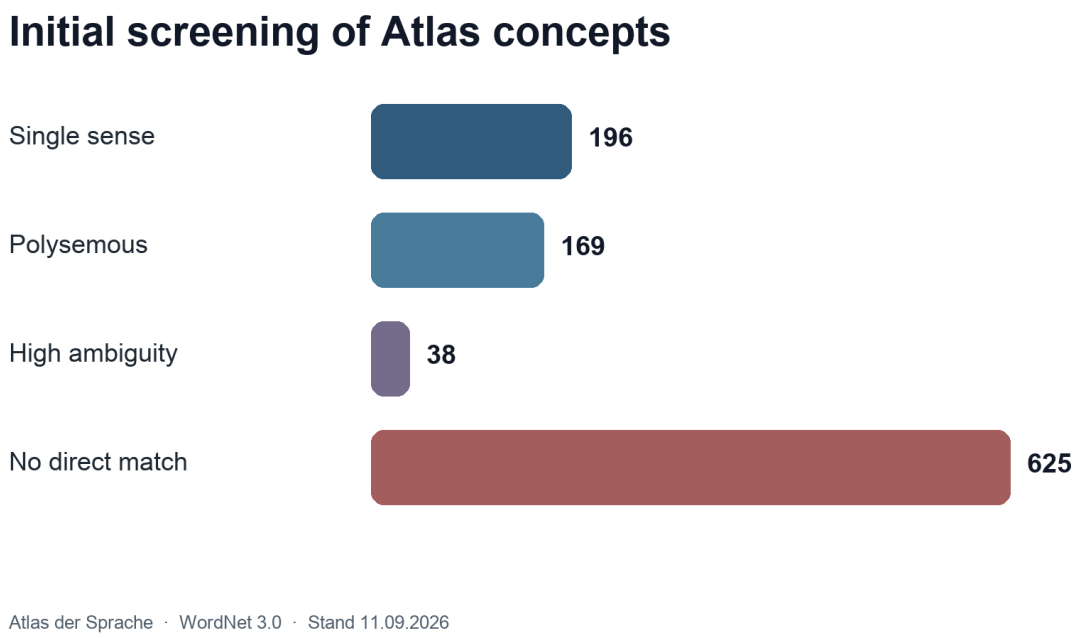

Figure 1 Initial screening of all 1,028 Atlas concepts

1 Research questions and scope boundary

The project distinguished three questions: whether an English Atlas lemma is present in WordNet; which concrete WordNet sense preserves the canonical Atlas definition; and whether an apparent absence reflects a genuine coverage gap or an unsuitable English Atlas lexicalization. These are interoperability questions, not claims of ontological identity.

WordNet groups English nouns, verbs, adjectives, and adverbs into synsets expressing distinct lexicalized concepts and connects them through lexical and conceptual relations (Princeton University, About WordNet). The Atlas contains many specialist, historical, rhetorical, linguistic, and media-theoretical

concepts whose granularity differs by design. A positive mapping therefore means that a synset is a defensible external anchor for the concept's semantic core. It does not mean that the two resources have identical definitions or modelling commitments.

2 Data and persistent identification

The study population consists of all 1,028 concepts in the canonical `concepts.jsonl` at the main revision stated above. Concept IDs, not strings, provide identity. The external authority is explicitly fixed to Princeton WordNet 3.0.

Sense keys were retained as the primary identifiers, with POS, synset offset, synset name, and gloss as review evidence. The official reference manual recommends sense keys for referring to concrete WordNet senses and notes that sense numbers and offsets vary between database versions (Princeton WordNet, `senseidx(5WN)`). WordNet 3.0 contains 117,659 synsets and 206,941 word–sense pairs across the four open lexical categories (Princeton WordNet, `wnstats(7WN)`); this scale does not imply exhaustive specialist-domain coverage.

3 Methods

3 1 Exhaustive first pass screening

All 1,028 English labels were normalized and queried in batches. Multiword expressions were tested in WordNet's underscore convention; regular case and morphological normalization was applied. The resulting strata were 196 `SINGLE_SENSE`, 169 `POLYSEMOUS`, 38 `HIGH_AMBIGUITY`, and 625 `NO_MATCH` cases.

Screening only measured the direct candidate space. It did not accept a mapping. This distinction follows WordNet's own architecture: word forms may have multiple senses, and a sense belongs to a specific synset with a shared gloss (Princeton WordNet, Glossary).

3 2 Manual sense review

For each reviewed concept, the canonical definition was compared with every lemma-exact WordNet gloss. Review dimensions were semantic core, domain, granularity, ontological type, part of speech, lexical form, and whether several synsets were needed. Decisions were recorded as exact/near, multiple-synset, or no adequate match. Negative decisions also received diagnostic labels identifying false friends, domain homographs, scope mismatches, or lexicalization issues.

The procedure deliberately allowed a documented near match when the same semantic core was preserved but one resource was narrower. It rejected a candidate when only a generic hypernym, a derived entity, an effect rather than a rhetorical strategy, or an unrelated domain sense remained.

3 3 Review 11

The final ten dossier cases produced nine exact/near mappings and one multiple-synset mapping. The selected concepts were Tautology, Circumlocution, Meiosis, Maxim, Cryptography, Layout, Shorthand, Typography, and Semantics; Facial expression required two senses because the Atlas integrates facial movement and expressed affect. No Review 11 case remained undecided.

3 4 Manual audit of 196 single sense hits

Every one of the 196 cases was inspected manually. A single WordNet candidate was accepted in 155 cases and rejected in 41. The rejected set comprised 21 false friends, ten external homographs lacking the relevant domain sense, and ten further scope or ontological-type mismatches. Thus, 20.9% of formally monosemous candidates were unsafe as automatic crosswalks.

This is a central methodological result: *monosemy within WordNet is not equivalent to unambiguous cross-resource identity*. Examples include musical *polyphony* versus narrative multivoicedness, the written *umlaut* mark versus the historical vowel-change process, and everyday *applicative* versus the grammatical construction.

3 5 Targeted QA of 625 `NO_MATCH` cases

The 625 cases were not exhaustively hand-reviewed. A high-recall procedure generated spelling, morphology, token-head, and definition-neighbor candidates. It returned 295 candidates, 277 of them triggered by some surface variant. Because generic heads produce many false positives, only all 30 high-information cases and four additional targeted compound probes were manually adjudicated.

The 34 decisions yielded eight robust findings: two Atlas concepts linked through WordNet's *antinomasia* spelling, three missed exact head senses (*grammatical number*, *lexical tone*, *numeral classifier*), one verb –event variant (*speaking in tongues*), one English normalization (*lexicalization*), and one two-sense result for *word coinage*. The remaining cases were partial heads, derivational neighbors, genuine compound gaps, or false neighbors.

The previously detected *Rückbildung → Reduction* problem was carried into the editorial gate as *Back-formation*. The new *Lexicization → Lexicalization* correction is handled in the same manner. Both were handled as explicit Atlas decisions. ATLAS 250 now uses Back-formation for Rückbildung and Lexicalization for Lexikalisierung while retaining the former English forms as provenance-preserving variants.

4 Results

The frozen row-level ledger contains 272 exact/near mappings, 26 multiple-synset mappings, 138 no-adequate-match decisions, and 592 explicitly unreviewed concepts. In total, 436 concepts have manual decisions and 298 have a defensible single- or multi-synset anchor. After rebase, editorial adjudication and controlled import, ATLAS 250 contains 342 WordNet sense-key crosswalks for 301 canonical concepts. Research class, package provenance and canonical import status remain separate data layers.

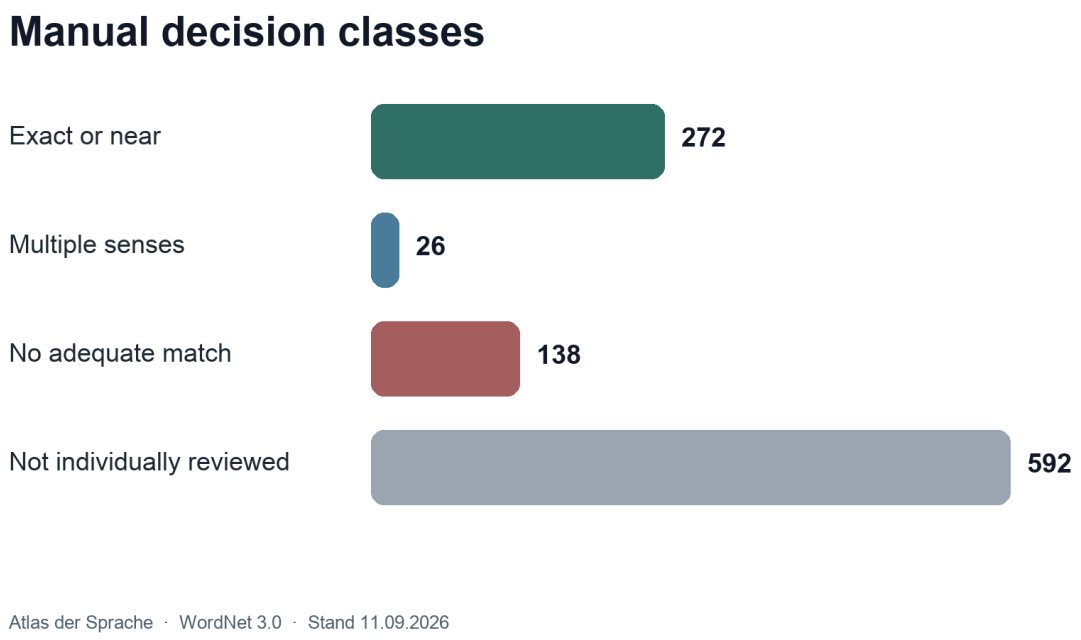

Figure 2 Frozen manual decision classes

The most reusable findings are:

- string presence, candidate count, and semantic adequacy are distinct measurements;
- specialist-domain shifts are a larger risk than simple spelling noise;
- compounds require definition-level review because their heads can be exact senses, hypernyms, or false friends;
- historical rhetoric and recent specialist linguistics form systematic WordNet coverage gaps;
- multiple-synset mappings can faithfully preserve integrated Atlas concepts rather than signal reviewer uncertainty.

5 Coverage gaps and Atlas lexicalization

The `NO_MATCH` population contains 326 specialist single tokens, 270 multiword expressions, and 29 hyphenated forms. The dominant gap types are historical specialist terms, compositional terms whose heads alone exist in WordNet, and granularity differences where the Atlas models a specific linguistic relation or practice.

Seventeen lexicalization or scope findings were transferred to the editorial layer. High-priority items include *Back-formation*, *Lexicalization*, *locutionary act*, *conversational repair*, *grammatical agreement*, *right dislocation*, and *applicative construction*. Domain qualification should be considered for rhetorical decorum, literary synaesthesia, rhetorical parabola, and narrative polyphony. Prolepsis and hyperbaton require possible concept-scope or subtype review.

Each finding received a separate editorial decision. Fourteen English labels were normalized or domain-qualified while their prior forms remained as variants. Common sense, Prolepsis and Hyperbaton were deferred because a label-only change would conceal unresolved concept history or scope. The earlier Rückbildung finding was additionally implemented as Back-formation. ATLAS 250 therefore contains fifteen accepted label corrections and three explicit deferrals.

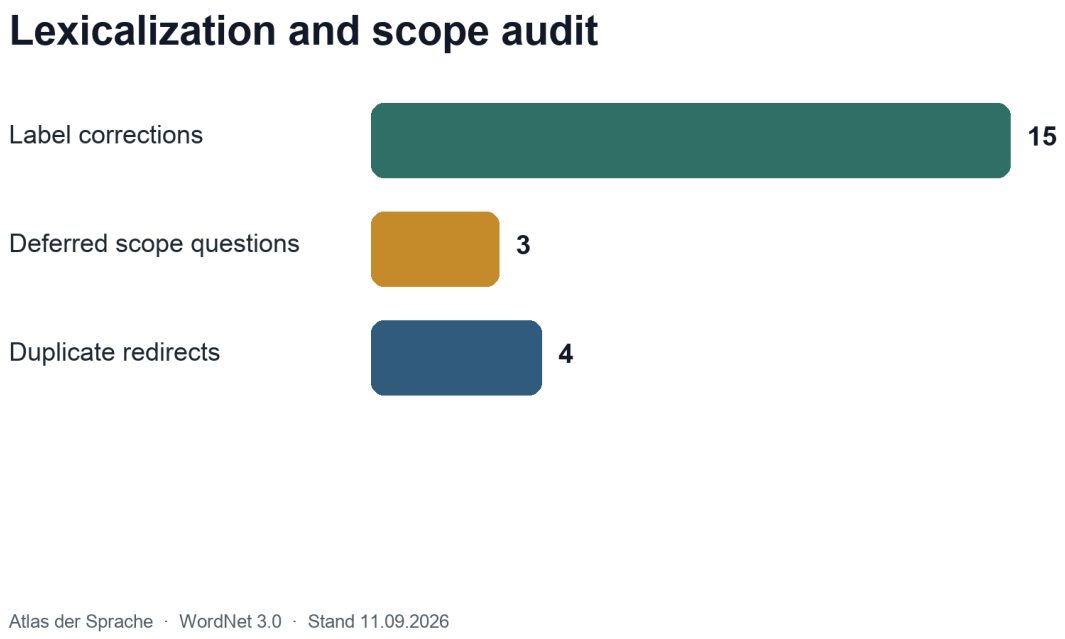

Figure 3 Editorial lexicalization decisions

6 Canonical main atlas gate and import

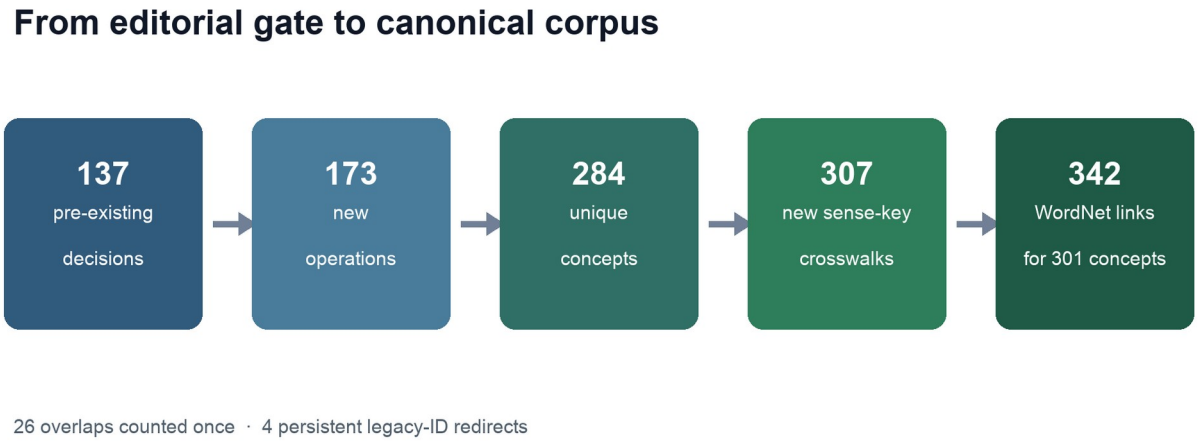

Figure 4 Consolidation and canonical import of Gate 284

The 137 pre-existing decisions remained identifiable as a separate input cohort during rebase. Together with 173 final-audit operations and 26 overlapping concepts, they formed a gate of 284 unique concepts represented by 320 retained provenance operations. Every concept decision is recorded in `WORDNET_CANONICAL_GATE_284_DECISIONS.json` with its source package, operation, sense key and target ID.

The accepted package `ATLAS-INBOX-WORDNET-GATE-284-ACCEPTED-V2` atomizes multi-sense mappings and deduplicates identical provenance, producing 307 new typed sense-key crosswalks. Acceptance relied on the preceding gloss, domain, granularity and ontological-type reviews rather than string equality or WordNet monosemy.

Four independently confirmed naming duplicates were merged in a migration separate from the additive crosswalk import: Anadiplosis into Anadiplose, Epanorthosis into Epanorthose, Symploke into Symploce, and Portmanteau into Mot-valise. Profiles, language forms, examples, evidence and external links moved to the survivor IDs. Six duplicate relation identities and two duplicate FrameNet identities were removed. Four persistent redirects retain the legacy IDs, and no ID was reused.

The canonical result contains 1,024 concepts, 3,427 directed relations, 1,739 evidence records, 872 typed external crosswalks, nine typological crosswalks and four redirects. The evidence total includes the 21 HTE links already present in the reconciled Atlas 8.15.6 base and is therefore not attributable to WordNet. The full migration sequence completed with `PRAGMA integrity_check = ok`, zero foreign-key errors and zero unresolved relations. Verified checkpoint `atlas-250-wordnet-gate-284-checkpoint` binds package SHA-256 `283b79b8e42335640d82b9d6b39fc27f27431cb2066ff0666af9fabfca7872d5`. The resulting export reproduces 342 WordNet links for 301 concepts.

7 Reproducibility, limitations, and conclusion

The freeze includes all screening batches, Reviews 1–11, the enriched 196-case single-sense audit, `NO_MATCH` typology and targeted decisions, a one-row-per-concept ledger, schema-valid proposal packages, and machine-readable plus narrative gate files. Assertions protect population sizes and gate arithmetic. Use of WordNet 3.0 is governed by Princeton's published license and attribution requirements (Princeton WordNet, License and Commercial Use). The project follows Princeton's requested scholarly citations to Miller (1995), Fellbaum (1998), and the online resource (Princeton WordNet, Citing WordNet).

Limitations remain: the study is version-specific; it is a curated expert review rather than an inter-annotator experiment; 591 `NO_MATCH` cases remain individually unreviewed; and exact/near mappings are not strict logical equivalence. The three deferred scope cases remain outside this release and must not be represented as silently resolved.

Within those boundaries, the research project and its canonical gate are complete. ATLAS 250 imports the reviewed sense anchors, preserves multi-sense integrations, records external coverage gaps and false friends, and keeps three unresolved scope questions visible. Public release 9.0.0 freezes this state; the next release remains blocked until the other research strands have reconciled the new canonical snapshot.

Literatur und References

- Fellbaum, Christiane (Hg.) (1998): *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press. ISBN 978-0-262-56116-7. Publisher record.
- Miller, George A. (1995): “WordNet: A Lexical Database for English.” *Communications of the ACM* 38(11), 39–41. Official citation guidance.
- Princeton University (2010–2026): *About WordNet*. Project description.
- Princeton University: *WordNet 3.0 Reference Manual*. Documentation index.
- Princeton University: *Sense Index File Format (`senseidx(5WN)`)* . Sense-key documentation.
- Princeton University: *WordNet 3.0 Database Statistics (`wnstats(7WN)`)* . Database statistics.
- Princeton University: *License and Commercial Use of WordNet*. WordNet 3.0 license.

Forschungsartefakte und Reproduzierbarkeit

- `SENSE_REVIEW_11_INPUT.json|md` und `SENSE_REVIEW_11_DECISIONS.md`
- `SINGLE_SENSE_AUDIT_INPUT.json|md`, `SINGLE_SENSE_AUDIT_DECISIONS.json|md`, `SINGLE_SENSE_AUDIT_SUMMARY.json`
- `NO_MATCH_TYPOLOGY.json|md`, `NO_MATCH_LEXICALIZATION_QA_CANDIDATES.json`, `NO_MATCH_LEXICALIZATION_QA_SHORTLIST.json|md`, `NO_MATCH_LEXICALIZATION_QA_DECISIONS.json|md`
- `FINAL_AUDIT_PACKAGE_STATUS.md`
- `CANONICAL_IMPORT_GATE.json|csv|md`
- `WORDNET_RESEARCH_LEDGER.csv|jsonl`, `LEDGER_SUMMARY.json`, `README.md`