Onomasiological Sense Alignment Across Dialect Dictionaries. A Taxonomy-Constrained LLM Classification

Nathalie Mederake, Nico Urbach, Hanna Fischer, Alfred Lameli
Research Center Deutscher Sprachatlas (DSA), Marburg University, Germany
{mederake|urbach|hanna.fischer|lameli}@uni-marburg.de

Abstract

We propose a taxonomy-guided approach to semantic alignment that assigns lexicographic senses to an onomasiological taxonomy derived from the Hallig–Wartburg/Post system. Using an LLM under strict taxonomic constraints, short and heterogeneous meaning descriptions are assigned to a common conceptual space. Evaluation against expert annotation shows that run-to-run model agreement ($\kappa = 0.73$) closely matches human agreement ($\kappa = 0.74$), with robustness at coarse taxonomic levels and predictable degradation at finer granularity. A qualitative network analysis demonstrates the resulting potential for cross-dictionary exploration of dialectal variation in semantics.

1 Introduction

Dialect dictionaries are a central resource for studying language variation and cultural knowledge (Reichmann, 2022), yet they remain notoriously difficult to compare or interlink. The core obstacle is not a lack of data but a lack of semantic interoperability: dialect dictionaries are typically organized semasiologically (i.e., by form), follow heterogeneous editorial conventions despite shared methodological foundations, and encode meaning in non-parallel ways. As a result, identical or closely related concepts are often realized through different expressions and scattered across isolated resources. This fragmentation severely constrains systematic cross-dictionary research on lexical variation.

An onomasiological perspective offers a fundamental alternative. Instead of starting from forms, it starts from meanings and groups words by the concepts they express. In German dialect lexicography, this perspective is well established through semantic taxonomies such as the Hallig-Wartburg system (HW; Hallig and von Wartburg, 1963) and its dialect-lexicographic adaptations. The HW taxonomy provides a conceptual framework that is independent of regional lexical realization and can, in principle (see Tittel et al., 2020), support concept-based access and comparison across dictionaries. In practice, however, onomasiological access is hard to apply consistently across resources: manual assignment is labor-intensive, semiautomatic methods struggle with sparse glosses, and fully automatic sense alignment techniques typically presuppose structurally homogeneous resources, richer definitions, or parallel sense inventories (Klee, 2024; Bieberstedt et al., 2024).

Large language models (LLMs) provide a qualitatively different option. Their ability to interpret underspecified meaning descriptions, integrate heterogeneous contextual cues, and follow explicit categorical constraints makes them promising candidates for taxonomy-bound semantic classification in dialect lexicography.[1] At the same time, deploying LLMs for this task raises a critical methodological question: can an LLM assign sense-levels to an onomasiological taxonomy reproducibly and lexicographically appropriately across divergent dictionaries, or do stochasticity and ambiguity undermine semantic coherence?

This paper addresses this question by treating taxonomy-bound classification as a practical route to sense-level assignment. We operationalize the HW taxonomy, as adapted for dialect lexicography by Post (1998), and employ it as an explicit conceptual framework. In its extended and partially restructured form (Post, 2026), Post's system of semantic categories serves as a controlled classificatory backbone, to which individual sense-levels are systematically assigned.

Empirically, we investigate the Hessen-Nassauisches Wörterbuch (HNWb) as a represen-

[1] In this paper, **meaning description** denotes meaning-related statements as they occur in dictionary texts. Methodologically, these statements are treated as **sense-levels**, that is, as levels of meaning differentiation within dictionary entries and used as operational units of analysis. **Semantic categories** are conceived as analytical classes used for taxonomic abstraction across individual meaning descriptions.

Proceedings of the Thirteenth Workshop on NLP for Similar Languages, Varieties and Dialects, pages 123–138
March 29, 2026.

tative and methodologically challenging dialect dictionary. We implement a reproducible processing pipeline that derives a fine-grained sense-level representation from dictionary encodings and performs LLM-based classification under strict taxonomic constraints. The approach is evaluated through expert-in-the-loop validation and repeated classification runs to quantify both agreement with expert judgments and run-to-run stability.

The paper thus makes three contributions: (i) we propose a taxonomy-guided method for sense-level alignment across dialect dictionaries; (ii) we provide an empirical evaluation of reproducibility via expert annotation and repeated runs; (iii) we illustrate, as a proof of concept, how sense-level assignments derived from the HNWb can be linked to corresponding categories in three other dialect dictionaries from major German dialect regions, laying the foundations for a systematic evaluation of cross-dictionary alignment in future work.

In doing so, we address three interrelated research questions:

RQ1: To what extent can LLM-based classification reliably and reproducibly assign sense-level entries from a single, structurally heterogeneous dialect dictionary to categories of the extended Post taxonomy?

RQ2: To what extent do LLM-generated assignments for the HNWb align with expert judgments under a defined evaluation scheme, and where do systematic deviations occur within the taxonomy?

RQ3: To what extent can a taxonomy-based, sense-level representation of the HNWb be used to link its senses to corresponding categories in other dialect dictionaries?

Beyond dialect lexicography, the proposed approach is, for example, applicable to domain-specific ontologies characterized by sparse and heterogeneous sense descriptions (e.g., historical, terminological, low resources languages).

2 Background

2.1 Semantic Interoperability in Dialect Lexicography

Prior research on sense alignment has proposed a variety of methods operating on both the expression and the content level (Wiegand and Gouws, 2014). Statistical measures used in information retrieval to assess the relevance of terms in documents within a document collection like TF-IDF (Term Frequency-Inverse Document Frequency) have been applied successfully to longer textual units (Lane and Dyshel, 2025). In lexicography-driven studies, information retrieval is producing particularly robust results for Realia, such as fruit or agricultural products (Burch and Rapp, 2007), but is ill-suited to short or fragmentary definitions. Content-based approaches have explored the comparison of conceptual information extracted from definitions, for example, through the use of hyperlemma lists (Fournier, 2003, 160–167). Sense-alignment work between GermaNet and DWDS has shown that parallel senses provide external evidence for the validity of sense distinctions and allow mutual enrichment of resources, whereas non-parallel cases point to entries in need of reconsideration (Henrich et al., 2014). Distributional methods relying on semantic proximity in vector spaces (Mikolov et al., 2013) offer another perspective, but they presuppose relatively balanced definitions and encounter difficulties when applied to dialectal or historical vocabulary.

More recently, cross-resource sense alignment has been framed as a supervised classification task over sense pairs, achieving improved performance through the combination of handcrafted features and representation learning (Ahmadi and McCrae, 2021). In a comparative case study of monolingual sense alignment, Salgado et al. (2020) highlight substantial challenges arising from inconsistent lexicographic criteria, heterogeneous sense inventories, and divergent wording practices in glosses. While their work demonstrates the practical value of sense alignment for semantic web publication and NLP applications, it also underscores that alignment quality is fundamentally constrained by the degree of semantic explicitness and structural comparability provided by the source dictionaries.

When applied to dialect lexicography, however, the assumptions underlying these approaches quickly reach their limits. Because many of these dictionaries are historical resources with inconsistent or minimal markup, the semantic cues they contain are not machine-readable. Even where technical aggregation infrastructures exist, such as in dictionary networks (e.g., `www.woerterbuchnetz.de`), the underlying semantic information remains incompatible (Klee, 2024). Interoperability at the level of individual lexicographic meaning descriptions, which is necessary for computational reuse in studies on language variation and change (Garrido García, 2021), lexicographic integration, or NLP applications, has therefore not kept pace with

the availability of digitized materials. This gap indicates that sense-level interoperability requires methods that can operate on sparse, heterogeneous definitions while remaining constrained by an explicit conceptual framework.

2.2 Semantic Taxonomy

Semantic taxonomies offer a potential remedy. Although originally developed within Romance studies, HW is a language-independent, onomasiological system designed to structure lexical meaning independently of lexical form. It functions as a controlled, hierarchical ontology and has been widely applied in historical lexicography and lexicological studies of Middle High German as well as medieval and early modern (up to the 16th century) French, Italian, Spanish, Gascon, and Occitan (cf. Tittel et al., 2020). It organizes concepts within a small number of highly abstract domains and a deeply articulated hierarchy whose ordering principle is intrinsic to the system (see Tittel et al., 2020 for a computational account). Concepts are positioned through their relations within this architecture rather than through assignment to mutually exclusive semantic fields.

Take the conceptual field *nature* as an example. In HW, the relevant concepts are embedded, for example, within the domain *A. The universe* and further structured along the path *I. The sky and the atmosphere* → *a) The sky and celestial bodies*, where elements such as *Nature*, *Wind*, and related phenomena appear as coordinated concepts within a system-internal hierarchy (original in French). Post's (2026) adaptation reorganizes this material for dialect lexicography. While largely preserving the conceptual inventory, Post restructures HW's hierarchical pathways into explicitly named and numerically indexed domains, for example, by differentiating *nature* into categories such as *Inanimate Nature* (Figure 1).

This restructuring transforms the HW system into an taxonomy designed for consistent lexicographic assignment. Post's (1998) modification is widely used in dialect dictionaries (e.g., Bickel, 2013; Breuer and Stöckle, 2023; Schwarz, 2022).[2]

By contrast, WordNet-style taxonomies are primarily synset-based and lexeme-driven, having been developed for contemporary English. Although they exhibit a hierarchical structure, their

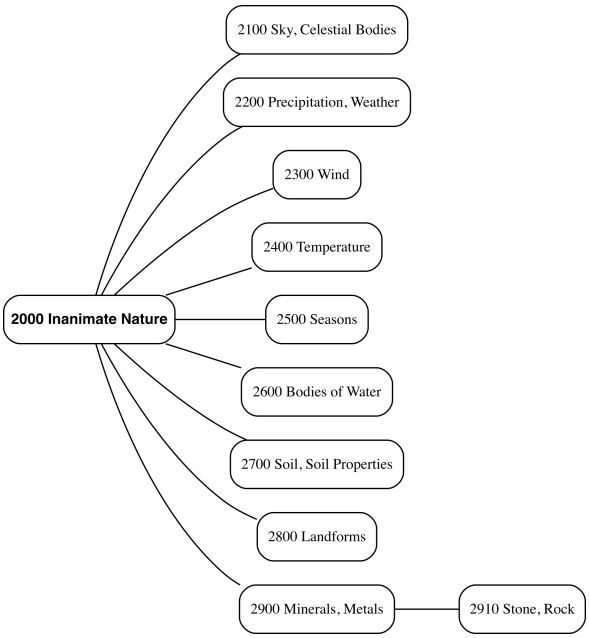

Figure 1: Extract from Post's (2026) modification of the HW taxonomy

organization is induced from the lexical inventory of a specific synchronic language stage. As a result, they are less suitable for lexical-semantic mapping in historical lexica or terminological resources, where conceptual structures may not align neatly with modern lexical units or where semantic relations diverge diachronically (see Trotter et al., 2016). The practical challenges of applying WordNet to historical lexical resources have received only limited attention in the literature (cf. Tittel, 2025). This structural limitation points to a fundamental difference: as a concept-driven and language-independent ontology, HW does not presuppose a particular synchronic lexicalization and therefore provides a more stable framework for diachronic and cross-linguistic mapping.

However, its computational application has remained experimental, for example, asking to what extent LLMs could effectively support the editorial process in a dictionary project (Raaf and Röhrer, 25.09.2025). Whether such a taxonomy can support automated large-scale alignment across dictionaries remains an open question.

3 Material and Methods

3.1 Dialect Data

This study takes the Hessen–Nassauisches Wörterbuch (HNWb) as its primary empirical object. While large-scale German dialect dictionaries are broadly comparable in their micro- and macrostruc-

[2]The version used in this paper follows the augmented taxonomy employed in the Fränkisches Wörterbuch (https://wbf.badw.de/wbf-digital/zur-dokumentation.html).

tural design (Lameli, 2021), they differ substantially from one another in editorial practice, temporal depth, and lexical coverage (Moulin, 2010; Lenz and Stöckle, 2020), limiting systematic cross-dictionary comparison. Focusing on the HNWb as a single case allows us to investigate taxonomy-based sense-level assignment under realistic lexicographic conditions. However, in order to illustrate the extensibility of the resulting sense-level representation, three additional dialect dictionaries are considered in a secondary role: the Mecklenburgisches Wörterbuch (MWb), the Pfälzisches Wörterbuch (PfWb), and the Schweizerisches Idiotikon (SchwId). Together, these resources cover major dialect regions from Low German (MWb) over Central German (HNWb, PfWb) to Upper German (SchwId) (Table 2).

The empirical analysis is restricted to the alphabetic range *N*, which is uniformly accessible across all four resources. While this restriction is motivated by data availability, it also offers methodological advantages: *N* spans a wide range of semantic domains (e.g., natural phenomena, temporal expressions, abstract operators), comprises both high-frequency function words and low-frequency content words, and exhibits substantial editorial heterogeneity across dictionaries. Furthermore, unlike other initial consonants, it does not exhibit onset variation (e.g., <f> vs. <v>, <t> vs <d>) and therefore provides a stable baseline for comparison. These properties make it a suitable test case for taxonomy-bound sense-level assignment.

Within the *N* range, the dictionaries differ considerably in size and microstructural granularity. The HNWb comprises 981 articles with 1,163 sense-levels. For the three additional dictionaries, the corresponding numbers are: MWb: 1492/1735; PfWb: 1502/1887; SchwId 3591/2982. Dictionary entries or single sense-levels lacking usable glosses—predominantly in SchwId—were excluded from the assignment.

3.2 Workflow

The methodological framework follows an onomasiological approach that aligns dictionary entries via the concepts they express. The HW taxonomy as extended by Post (2026) serves as a controlled conceptual foundation. Each sense-level is mapped to a four-digit taxonomy code, positioning it within a hierarchical domain and thereby enabling nested evaluation across different levels of conceptual granularity (e.g., code 2000 "Inanimate Nature", extended by 2200 "Precipitation and Weather"; see Figure 1). The definition of sense-levels is based on the Standard German definitions provided in the dictionaries, rather than on dialectal keywords. Examples of dialectal usage, which typically take the form of multiword expressions, were explicitly excluded from the analysis because they often refer to multiple lexical items.

The LLM classification follows a detailed workflow (Figure 2) outlined in the following sections.

Preprocessing The HNWb entries were brought to a uniform level of annotation with regard to meaning specification using lightweight markers inspired by TUSTEP conventions (e.g., $\#marker+ \ldots \#marker-$). The resulting information was stored in a JSON schema that explicitly separates (a) lexical form, (b) grammatical metadata, (c) sense-levels, and (d) auxiliary gloss material (e.g., Latin equivalents). This schema accommodates heterogeneous metadata while enabling systematic extraction of sense-levels, batch processing, and reproducible interaction with the LLM-based classification workflow. Structural inconsistencies and reference entries were resolved manually according to pre-defined annotation rules. The same preprocessing framework was subsequently applied to the additional dialect dictionaries used for illustrative cross-dictionary linking, ensuring structural compatibility at the sense-level.

Prompt Engineering Semantic classification was performed using the OpenAI model GPT-5, accessed via the KISSKI interface.[3] We employed an in-context learning prompt (Figure 10) comprising five carefully selected few-shot examples. These examples were designed to enforce strict taxonomic constraints while minimizing over-interpretation of dialectal form variation (e.g., phonological and morphological variation). Dictionary entries were processed in batches of approximately 8,000–9,000 tokens. Using this procedure, all entries in the alphabetic range *N* of the HNWb were distributed across 14 subsets, each containing all information required for classification.

[3] AI Service Centre for Sensitive and Critical Infrastructure; see https://kisski.gwdg.de for details. The portal provides a simplified API interface. Any associated costs were billed directly through the University of Marburg, which was essential given the absence of dedicated project funding. Model outputs were generated using default sampling parameters (temperature = 0.5, top-p = 0.5, no frequency or presence penalty). This configuration corresponds to the KISSKI default and is reported to OpenAI.

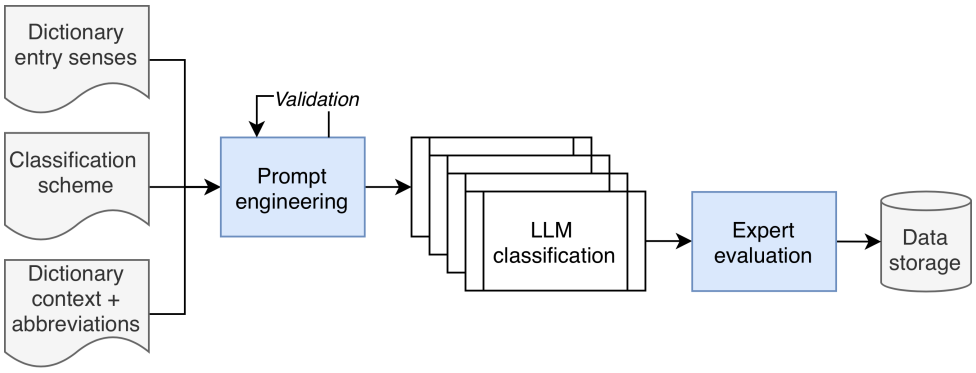

Figure 2: Workflow for LLM sense-level alignment.

Entries that did not contain semantic information and merely referred to other entries–either within the same batch or outside the *N* range–were labeled as "out of index" and excluded from further evaluation. Since such entries point to other lemmas via explicit cross-references, assigning their meanings constitutes a straightforward rule-based task. In addition, almost all categorizations of these reference entries produced by the LLM were correct.

To ensure consistent and context-aware classification, each batch–comprising the lemma entry, variants, grammatical information, and sense-levels–was paired with a comprehensive context file. This file included abbreviation lists, editorial metadata, dialect background information, and the full extended Post taxonomy (approximately 10,700 tokens[4]).

Classification The automated semantic classification was conducted in three stages.

Stage 1 (primary assignment): The LLM-based classifier was first applied to the HNWb in order to generate taxonomy-based sense-level assignments under controlled prompt settings (1st run). The classification was carried out in multiple iterative calls, with a practical upper limit of approx. 20,000 tokens per call.

Stage 2 (reproducibility assessment). To assess run-to-run reliability, the HNWb was classified a second time under identical prompt settings and experimental conditions (2nd run). This 2nd run served to quantify the stability and reproducibility of the automated sense-level assignment. The results of the two runs were evaluated on the basis of expert-validated sense-level assignments.

Stage 3 (illustrative extension). In a final step (3rd run), the same classification procedure was applied to two additional dialect dictionaries (MWb, SchwId). These assignments were not subjected to the same reproducibility or expert-based evaluation but served to demonstrate how sense-level representations derived for the HNWb can be linked to, a broader dialect-lexicographic context. Additionally, we used the available sense-level assignments from the PfWb (see Post, 1998).

Expert Evaluation To assess the quality and reliability of the LLM-based semantic classification, the model-generated sense-level assignments of the HNWb were evaluated by two expert lexicographers at two distinct points in the workflow.

Evaluation 1 (quality check for prompt design): A preliminary quality check was conducted on a random sample of approx. 100–120 sense-level assignments per dictionary (Validation during prompt engineering in Figure 2). The purpose of this check was to verify whether the prompt design and model behavior met the requirements for large-scale application. Each assignment was manually inspected and judged dichotomously (correct vs. incorrect) by the expert lexicographers working independently. An acceptance threshold of 80% correct assignments was defined ex ante as the minimum level required to justify full deployment. This value was chosen conservatively based on a pilot evaluation of the random sample and serves as an operational decision point within the pretesting phase. If the threshold was not met, the prompt was revised and the evaluation repeated.

Evaluation 2 (quality check after classification): A second expert evaluation was carried out after completion of the 1st run, focusing exclusively on the 997 assigned sense-level entries of range *N* in the HNWb. All assignments were independently corrected by the same two lexicographers. This led to two individual classifications which were

[4]Tokenization as determined by the `tiktoken` 0.12.0 Python library.

statistically evaluated in terms of reliability. To this end, a three-level assessment scheme was applied: (1) *Correct*, indicating an exact match between the model-generated and expert-assigned category; (2) *Almost correct*, indicating assignment to a semantically adjacent node within the correct higher-level category; and (3) *False*, covering all remaining cases. The individual classifications were then compared and any disagreements were resolved to arrive at a consensual final assignment. The finalized expert assignments derived from this evaluation subsequently served as the baseline for the automated evaluation of the 2nd run.

4 Results

4.1 Sense-Level Assignment

Human Baseline To establish a reference point for the level of agreement achievable under Post's semantic taxonomy, inter-rater agreement between the two expert lexicographers was calculated for all entries in the HNWb *N* range. Agreement was quantified using unweighted Cohen's κ, yielding a value of $\kappa = 0.743$. This level reflects the intrinsic complexity of the classification task. Post's taxonomy is hierarchically structured, and category boundaries are not always sharply defined. Consequently, variability in expert judgments should be interpreted as an effect of semantic granularity and ambiguity rather than as annotator error.

Run-to-run Reliability To assess the internal consistency of the LLM-based classification, the two independent classification runs on the HNWb were compared using identical prompts and experimental conditions. The expert-validated sense-level assignments served as the reference standard. Agreement between the 1st and 2nd run was substantial ($\kappa = 0.733$), closely approximating the human baseline. In absolute terms, the proportion of strictly correct assignments decreased from 86.9% (875 cases) in the 1st run to 81.0% (812 cases) in the 2nd run. This difference is statistically significant ($\chi^2(1) = 14.42$, $p < .001$), although the associated effect size was small ($\phi = .08$).

Stability Beyond aggregate agreement, the item-level overlap between the two runs was examined (Table 1). Of the assignments classified as strictly correct in the 1st run, 89.40% (776 cases) retained this status in the 2nd run. By contrast, assignments classified as almost correct showed lower retention (60.34%, 35 cases). Incorrect classifications were retained in 74.64% (53 cases). Across all items, roughly 25% of sense-level assignments changed their evaluation level (correct / almost correct / false) between runs (Figure 3).

Evaluation	1st run	2nd run	Overlap
Correct	868	811	776 (89.40%)
Almost correct	58	68	35 (60.34%)
False	71	111	53 (74.64%)

Table 1: Overlap of sense-level alignments between 1st and 2nd run.

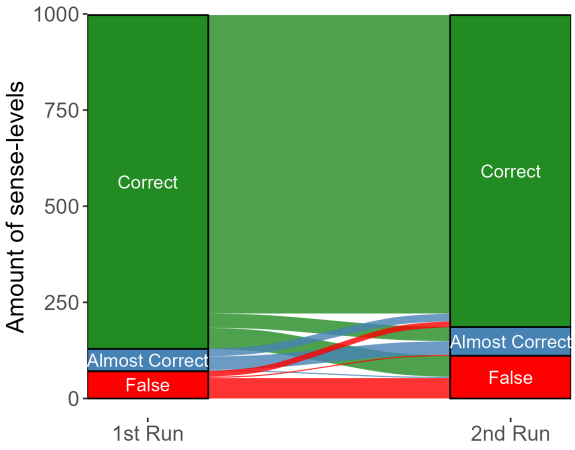

Figure 3: Item overlap of sense-level alignments between 1st and 2nd run.

Taxonomic Granularity Classification accuracy varies systematically with taxonomic depth. When evaluated at higher, more abstract levels of Post's taxonomy, assignments show higher accuracy and greater stability. Accuracy decreases as categories become more fine-grained (Figure 4), with increased variability across repeated runs. Error bars indicate 95% Wilson confidence intervals for binomial proportions.

4.2 Semantic Inspection

Taxonomy-Sensitive Stability Beyond aggregate agreement scores, run-to-run stability varies systematically across semantic regions of the taxonomy. As shown in Figure 6 and Figure 7, sense-level entries of the HNWb differ markedly in their propensity to change categories between the two runs. Highly referential domains (e.g., 3000 "Plants and fruits") show consistently low

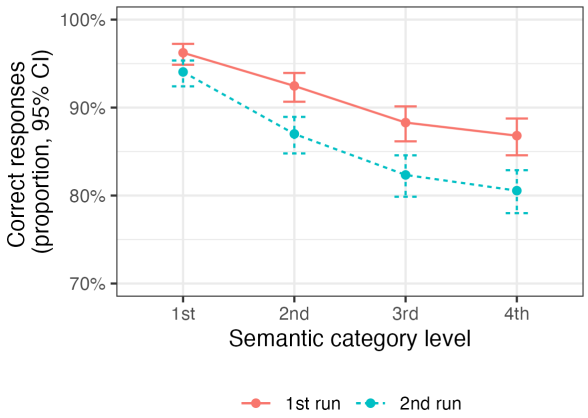

Figure 4: Classification accuracy across hierarchical semantic levels; error bars indicate 95% Wilson confidence intervals for binomial proportions.

change rates and high proportions of strictly correct assignments (see Figure 8), whereas, for example, abstract or evaluative categories exhibit substantially higher variability. There are also collective categories that have no semantic relevance and therefore lead to misclassifications, such as the high-level category 1000 "Special Words" that, among others, contain temporal and spatial adverbs which—semantically—could be attributed to 9500 "Space" or 9600 "Time".

A further, qualitatively distinct source of misclassification arises in normatively charged areas of the taxonomy. Meanings referring to human behavior or personal traits are repeatedly assigned to evaluative categories such as 6116 "improper behaviour" or 6510 "moral judgment (adjectival)" (Figure 8), even when no moral judgment is encoded in the meaning description. Importantly, these missassignments are systematic rather than random and suggest that the model tends to activate evaluative supercategories when processing human-related semantics. The phenomenon points to an interaction between the taxonomy's normative structuring and the model's tendency to infer implicit evaluation, rather than to isolated classification errors.

Semantic Network We further illustrate how taxonomy-based, sense-level assignment derived from the HNWb can restructure access to dialect dictionary data beyond traditional alphabetical lookup. While the distributional analysis establishes robustness at the level of aggregated categories, it does not show how individual sense-level entries relate to one another within the shared conceptual space. To make this relational structure visible, we include sense-level assignments from three additional dialect dictionaries obtained in the 3rd run and project all classified senses into a common taxonomic space for exploratory purposes. This space is represented as a semantic network of cross-dictionary relations. For illustration, we focus on sense-level entries across all subcategories under node 2000 "Inanimate Nature" (Figure 9), a region of the taxonomy that shows comparatively low volatility across runs (Figure 6). Within this network, the categories of Post's taxonomy delineate conceptual regions, within which individual sense-level entries are positioned as nodes.

Mixed colors within a node indicate categories with strong cross-dictionary representation. Take category 2100 "Sky, Celestial Bodies" as an example, where the concept 'new moon' forms a tightly connected semantic neighborhood (Figure 5). Across dictionaries, formally divergent but etymologically related designations converge within this category. Entries such as *Neumond* (hnwb_N00568), *Neu-mond* (pfwb_PN01033), *Niman* (mwb_N01074), and *Nūū(w)mān* (schwi_159172) share a common etymon and are thus connected as cognate realizations of the same basic concept. In addition, synonymous expressions based on alternative lexical motivations, such as *Neulicht* (hnwb_N00565), *Nūū(w)* (schwi_159114), and *Neu-schein* (pfwb_PN01081), are integrated into the same semantic region.

Beyond naming variants, the network also captures systematically related concepts that extend the semantic field of *new moon*. These include oppositional or contrastive relations, as in *Nachthelling* 'moonlight' (pfwb_PN00239), as well as derivational extensions such as the adjective *nūū(w)mānlich* 'taking place at the time of the new moon' (schwi_159173). By bringing together cognate forms, synonyms, and semantically related derivations from multiple dictionaries within a single taxonomic category, the network reveals conceptual relationships that would be difficult to detect systematically through alphabetical access.

The resulting network shows that sense-level entries from the HNWb consistently co-occur with semantically corresponding entries from the additional dictionaries within the same taxonomic regions. Formally divergent lexicalizations, including cognate forms, synonymous expressions, and derivationally related senses, cluster in coherent conceptual neighborhoods, indicating that the taxonomy-based sense-level assignments support meaningful cross-dictionary linking.

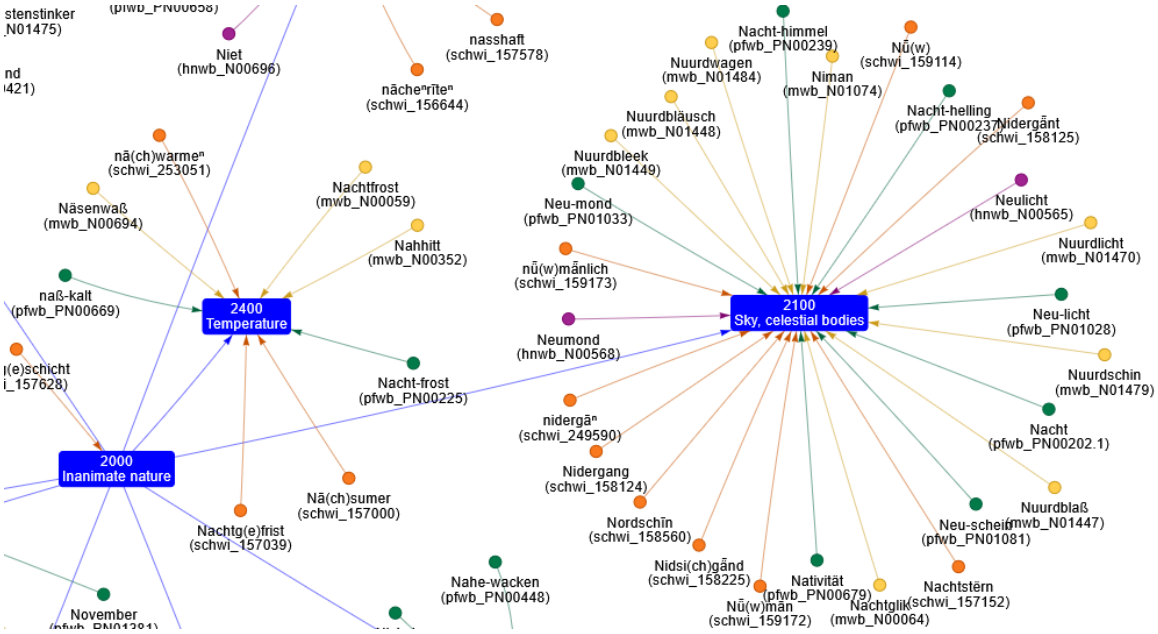

Figure 5: Semantic network of sense-levels for node 2000 "Inanimate Nature" of the taxonomy following Post (2026) (extraction of Figure 9); colors according to dictionaries, labels with lemma and ID of sense-levels.

5 Discussion

Reliability Relative to Human Agreement Addressing RQ1 and RQ2, the results show that taxonomy-bound LLM classification achieves a level of reliability comparable to expert annotation. Human inter-rater agreement ($\kappa = 0.743$) and run-to-run model agreement ($\kappa = 0.733$) fall into a similar range, indicating that a substantial share of classification variability reflects genuine semantic ambiguity rather than model instability (see Artstein and Poesio, 2008; Passonneau and Carpenter, 2014). Given the hierarchical structure of Post's taxonomy and the often underspecified or context-dependent nature of meaning descriptions in dialect dictionaries, perfect agreement cannot be expected. Convergence between human and model agreement therefore suggests that the LLM operates within the bounds of lexicographic interpretative practice, rather than introducing qualitatively different error patterns. From this perspective, volatility becomes an informative signal: it highlights regions where the taxonomy enforces discrete choices on meanings that are inherently perspectival or multi-dimensional, and thus marks the practical limits of fine-grained, deterministic classification.

Sources of Instability and Semantic Ambiguity Item-level analysis reveals that instability is concentrated in borderline cases. Assignments classified as almost correct show lower retention across runs than strictly correct or incorrect assignments, indicating competition between semantically adjacent taxonomic categories. Incorrect classifications, by contrast, tend to remain stable once a decision falls outside the correct taxonomic region. This asymmetry suggests that variability is driven by semantic underspecification rather than random noise.

Overall, approx. 25% of sense-level assignments change their evaluation-status across runs. Rather than undermining reliability, this volatility provides an empirical estimate of the proportion of cases for which fine-grained taxonomic assignment is inherently uncertain. This is not solely due to the classification, but also to the taxonomy, which in some cases contains vague categories or collective categories (e.g., 9214 "other"). At the same time, the comparison of the 1st and the 2nd run reveals a theoretical upper bound for semantic assignment accuracy. Although the classifications of the runs overlap substantially, they differ in 35 strictly correct assignments (see Table1). Aggregating all correct assignments in either run would raise accuracy to 90.57%. This constitutes a best-case estimate conditioned on the two runs. Given the probabilistic nature of the classification and the observed volatility, deterministic aggregation is not feasible in practice. However, the complementarity across runs suggests that controlled resampling may increase coverage, an issue left for future work.

Furthermore, the strong dependence of accuracy

on taxonomic depth highlights a central trade-off: coarse-grained categories (1st or 2nd level in the taxonomy) are assigned robustly and reproducibly, whereas fine-grained distinctions (3rd or 4th level) show increased variability. This pattern mirrors human annotation behavior under hierarchical taxonomies and indicates that taxonomy-based alignment is suited for applications that operate at intermediate or higher levels of abstraction, while finer distinctions should be interpreted probabilistically.

Qualitative Sense-Level Assignment and Cross-Dictionary Linking Addressing RQ3, qualitative inspection of taxonomy-based semantic networks demonstrates how a sense-level representation derived from the HNWb can be extended beyond alphabetical access and linked to other dialect dictionaries (see 190 Klosa and Müller-Spitzer, 2016; Moulin, 2010). Projecting sense-level entries from multiple resources into a shared taxonomic space, reveals coherent cross-dictionary clustering within individual categories. Formally divergent lexicalizations (Niebaum, 1986)—including cognate forms, synonymous expressions, and derivational extensions—consistently co-occur within the same conceptual regions, indicating successful taxonomy-based linking at the sense-level.

Crucially, the graph-based representation renders explicit conceptual relations that would not be readily observable through linear, alphabetically organized dictionary access. For example, lexical items such as *Neulicht* 'new moon' (hnwb_N00565; cf. Figure 5), corresponding to Standard German *Neumond*—with *Neulicht* literally meaning 'new light'—and *Nachthimmel* 'night sky' (pfwb_PN00239) would not ordinarily be associated on the basis of alphabetical ordering or surface-level semantic similarity. Within the taxonomic network, however, their relation becomes apparent through shared higher-level conceptual nodes, thereby exposing latent semantic connections and yielding interpretive added value.

Beyond facilitating cross-dictionary exploration, this perspective also highlights a structural property of parts of dialect lexicography: some dialect dictionaries document the differential lexicon relative to the standard language more systematically than basic conceptual inventories. Apparent gaps in basic concepts therefore often reflect editorial practice rather than conceptual or lexical absence. Taxonomy-based alignment makes such asymmetries explicit and enables comparative analysis of how semantic domains are lexically populated across regions. At the same time, recurrent misclassifications and systematic category overextensions observed in the LLM-based assignments provide empirical feedback for refining the taxonomy itself, identifying nodes that are overly abstract, normatively overloaded, or insufficiently discriminative. In this way, taxonomy-guided sense-level assignment not only supports linking but also serves as a diagnostic lens on both lexicographic practice and the taxonomy.

Scope and Limits of Taxonomy-Based Alignment Taxonomy-guided sense-level assignment does not, by itself, resolve the full problem of semantic mapping. Assigning multiple sense-levels to the same Post category (e.g., 2100 "Sky, Celestial Bodies") situates them within a shared conceptual space, but it does not entail semantic equivalence at finer levels of differentiation. Within this framework, the Post taxonomy functions as a controlled conceptual backbone that constrains the semantic search space and enables first-order alignment across heterogeneous sense descriptions. Further differentiation within a category would require additional analytical layers, such as relation typing or sub-clustering at the node level. Taxonomy-based classification should therefore be understood as a foundational step toward sense-level interoperability: it provides a stable, hierarchically organized structure that conditions—but does not substitute for—more fine-grained semantic analysis. At the same time, grouping semantically related concepts supports resource-efficient workflows, particularly important for LLM-assisted approaches to identifying semantic relations between dictionary entries.

6 Conclusion

This paper introduced a reproducible workflow for taxonomy-based, sense-level assignment using the HNWb as a case study. Combining a semantic taxonomy with LLM-supported classification and expert evaluation, we show that meaning descriptions are reliably assigned to common semantic categories, with robust reproducibility across sense-levels. The resulting representation supports structured semantic access and enables linking to other dialect dictionaries, establishing taxonomy-guided LLM classification as a foundation for sense-level interoperability in dialect lexicography and related low-resource NLP settings.

Limitations

Several limitations of the present study should be acknowledged. First, the empirical evaluation is restricted to the alphabetic range *N*, which is the only segment uniformly available across all four dictionaries. Although this range spans multiple semantic domains and provides a meaningful stress test for interoperability, it does not support claims regarding full-dictionary coverage or domain-complete semantic analysis. Extending the approach to additional alphabetic ranges remains a task for future work.

Second, the proposed approach depends on the availability and quality of explicit meaning descriptions. Dictionary entries and single sense-levels lacking usable glosses were excluded from classification, and deviations from Standard German in the metalanguage (e.g., mixed dialect–standard formulations or Latin quotations) required additional preprocessing. Dictionaries characterized by minimal, implicit, or highly heterogeneous meaning descriptions may therefore necessitate further normalization efforts.

Third, classification quality is influenced by prompt configuration, contextual information, and the specific LLM version and interface employed. Although the workflow is reproducible, alternative model settings or future model updates may require re-tuning of prompts and contextual parameters. Moreover, while expert-in-the-loop validation ensures high-quality assignments, it introduces manual effort that may limit scalability in fully automated scenarios.

Finally, the present study focuses on establishing methodological feasibility and semantic plausibility rather than on evaluating downstream applications. Tasks such as systematic variation of batch size and order, controlled resampling of sense-level alignments, and large-scale clustering or similarity measurement constitute important directions for future research. While the study covers a wide range of structurally diverse dialects, its findings are necessarily limited to variation within German. Nonetheless, the proposed methodology is not language-specific and can, in principle, be extended to other languages and dialect continua.

Acknowledgments

This work is funded by the Academy of Sciences and Literature Mainz (Grant REDE 0404). We would like to thank three anonymous reviewers for their valuable comments and discussion. We would also like to thank Rudolf Post for sharing the sense-level assignments of the PfWb.

References

Sina Ahmadi and John P. McCrae. 2021. Monolingual word sense alignment as a classification problem. In *Proceedings of the 11th Global Wordnet Conference*, pages 73–80, University of South Africa (UNISA). Global Wordnet Association.

Ron Artstein and Massimo Poesio. 2008. Survey article: Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Hans Bickel. 2013. Fortschreitende Digitalisierung. Neue Zugriffe auf das Idiotikon. In *150 Jahre Schweizerisches Idiotikon: Beiträge zum Jubiläumskolloquium in Bern, 15. Juni 2012*, pages 121–134, Bern. Schweizerische Akademie der Geistes- und Sozialwissenschaften.

Andreas Bieberstedt, Nico Förster, Petra Himstedt-Vaid, and Christoph Schmitt. 2024. Vom gedruckten Wörterbuch zur virtuellen Forschungsumgebung: Digitale Vernetzungsszenarien dialektaler Großwörterbücher am Beispiel des Mecklenburgischen Wörterbuchs. In Antje Dammel and Markus Denkler, editors, *Grosslandschaftliche Dialektwörterbücher zwischen Linguistik und Landeskunde*, pages 17–55. Böhlau Verlag, Wien, Köln.

Ludwig M. Breuer and Philipp Stöckle. 2023. Informationssystem Österreich - Zugriff und Vernetzungsmöglichkeiten: Version 2 (22.06.2023, 16:23).

Thomas Burch and Andrea Rapp. 2007. Das Wörterbuchnetz: Verfahren – Methoden – Perspektiven. In Daniel Burckhardt, Rüdiger Hohls, and Claudia Prinz, editors, *Geschichte im Netz: Praxis Chancen Visionen. Beiträge der Tagung .hist 2006*, pages 607–627. Humboldt Universität zu Berlin, Berlin.

Johannes Fournier. 2003. Vorüberlegungen zum Aufbau eines Verbundes von Dialektwörterbüchern. *Zeitschrift für Dialektologie und Linguistik*, (70):155–176.

C. Garrido García. 2021. *A structural analysis of dictionaries as semantic networks*. Ph.D. thesis, Universidad de Chile.

Rudolf Hallig and Walther von Wartburg. 1963. *Begriffssystem als Grundlage für die Lexikographie: Versuch eines Ordnungsschemas*. Veröffentlichungen des Instituts für Romanische Sprachwissenschaft. Akademie-Verlag, Berlin.

Verena Henrich, Erhard W. Hinrichs, and Reinhild Barkey. 2014. Aligning word senses in germanet and the dwds dictionary of the german language. In *Global WordNet Conference*.

Anne Klee. 2024. Vernetzungsstrategien zwischen Dialektwörterbüchern - am Beispiel des Trierer Wörterbuchnetzes. In Antje Dammel and Markus Denkler, editors, *Grosslandschaftliche Dialektwörterbücher zwischen Linguistik und Landeskunde*, pages 113–131. Böhlau Verlag, Wien, Köln.

Annette Klosa and Carolin Müller-Spitzer, editors. 2016. *Internetlexikografie*. De Gruyter, Berlin, Boston.

Alfred Lameli. 2021. Dialektwörterbücher zwischen Web 0.0 und Web 3.0. In *Das Sudetendeutsche Wörterbuch: Bilanzen und Perspektiven*, pages 45–70. Frank & Timme.

Hobson Lane and Maria Dyshel. 2025. *Natural language processing in action*, second edition edition. Manning Publications, Shelter Island, NY.

Alexandra N. Lenz and Philipp Stöckle. 2020. *Germanistische Dialektlexikographie zu Beginn des 21. Jahrhunderts*. Steiner, Stuttgart.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *Preprint*, arXiv:1301.3781.

Claudine Moulin. 2010. 33. Dialect dictionaries - traditional and modern. In Peter Auer and Jürgen Erich Schmidt, editors, *Language and Space. An International Handbook of Linguistic Variation. Volume 1 Theories and Methods*, pages 592–612. Walter de Gruyter, Berlin, New York.

Hermann Niebaum. 1986. Lemma und Interpretament: Zur Problematik der Artikelgestaltung in Dialektwörterbüchern. In Hans Friebertshäuser and Heinrich J. Dingeldein, editors, *Lexikographie der Dialekte*, pages 125–144. De Gruyter, Berlin.

Rebecca J. Passonneau and Bob Carpenter. 2014. The benefits of a model of annotation. *Transactions of the Association for Computational Linguistics*, 2:311–326.

Rudolf Post. 1998. Möglichkeiten der elektronischen Strukturierung, Vernetzung und Verfügbarmachung von lexikographischen Daten bei der Arbeit am Pfälzischen Wörterbuch. In Rudolf Grosse, editor, *Bedeutungserfassung und Bedeutungsbeschreibung in historischen und dialektologischen Wörterbüchern: Beiträge zu einer Arbeitstagung der deutschsprachigen Wörterbücher, Projekte an Akademien und Universitäten vom 7. bis 9. März 1996 anläßlich des 150jährigen Jubiläums der Sächsischen Akademie der Wissenschaften zu Leipzig*, pages 211–220. Hirzel, Stuttgart.

Rudolf Post. 2026. Postsche Sachkategorien - CSV. LinguRep.

Manuel Raaf and Ines Röhrer. 25.09.2025. Semantische Klassifikation lexikographischer Inhalte mithilfe künstlicher Intelligenz neu denken? Ergebnisse einer Studie zur Erweiterung des Bayerischen Wörterbuchs um Ontologie mithilfe von LLMs. FORGE 2025 - Daten neu denken (FORGE), Rostock.

Oskar Reichmann. 2022. Dimensionen der Wortbedeutung und historische Lexikographie. In Gerhard Diehl and Volker Harm, editors, *Historische Lexikographie des Deutschen. Perspektiven eines Forschungsfeldes im digitalen Zeitalter*, pages 229–254. De Gruyter, Berlin, Boston.

Ana Salgado, Sina Ahmadi, Alberto Simões, John P. McCrae, and Rute Costa. 2020. Challenges of word sense alignment: Portuguese language resources. In *Workshop on Linked Data in Linguistics*.

Brigitte Schwarz. 2022. *Das dialektologische Informationssystem von Bayerisch-Schwaben: Dokumentation und mögliche Präsentation von Sprachdaten mit Multimedia im Internet*. Ph.D. thesis, Universität Augsburg, Stuttgart.

Sabine Tittel. 2025. Historisierte Ontologien für Linguistic Linked Open Data-Ressourcen des Mittelalters. In *Das Mittelalter. Perspektiven mediävistischer Forschung*. Heidelberg University Publishing.

Sabine Tittel, Frances Gillis-Webber, and Alessandro A. Nannini. 2020. Towards an ontology based on Hallig-Wartburg's Begriffssystem for historical linguistic linked data. In *Proceedings of the 7th Workshop on Linked Data in Linguistics (LDL-2020)*, pages 1–10, Marseille, France. European Language Resources Association.

David Trotter, Andrea Bozzi, and Cédrick Fairon. 2016. Un'ontologia per il ditmao (dictionnaire des termes médico-botaniques de l'ancien occitan). In *Actes du XXVIIe Congrés international de linguistique et de philologie romanes (Nancy, 15-20 juillet 2013). Section 16: Projets en cours; ressources et outils nouveaux*. ATILF, Nancy.

Herbert Ernst Wiegand and Rufus H. Gouws. 2014. Access structures in printed dictionaries. In Rufus H. Gouws, Ulrich Heid, Wolfgang Schweickard, and Herbert Ernst Wiegand, editors, *Dictionaries: An International Encyclopedia of Lexicography*, volume 5.4 of *Handbooks of Linguistics and Communication Science*, pages 110–148. De Gruyter Mouton, Wien, Köln.

A Appendix

	MWb	HNWb	PfWb	SchwId
Project launch	1926	1911	1912	1862
Publication	1937–1998	Since 1927	1965–1997	Since 1881
Dialect area	Low German	Central German	Central German	Upper German
Survey period	13th–20th cent.	1912–1934	ca. 1900–1950	13th–21st cent.
Hist. references	Yes	No	No	Yes
Headwords	Dialect	Standard German	Standard German	Dialect

Table 2: Key characteristics of the dialect dictionaries analyzed: Mecklenburgisches Wörterbuch (MWb), Hessen-Nassauisches Wörterbuch (HNWb), Pfälzisches Wörterbuch (PfWb) and Schweizerisches Idiotikon (SchwId).

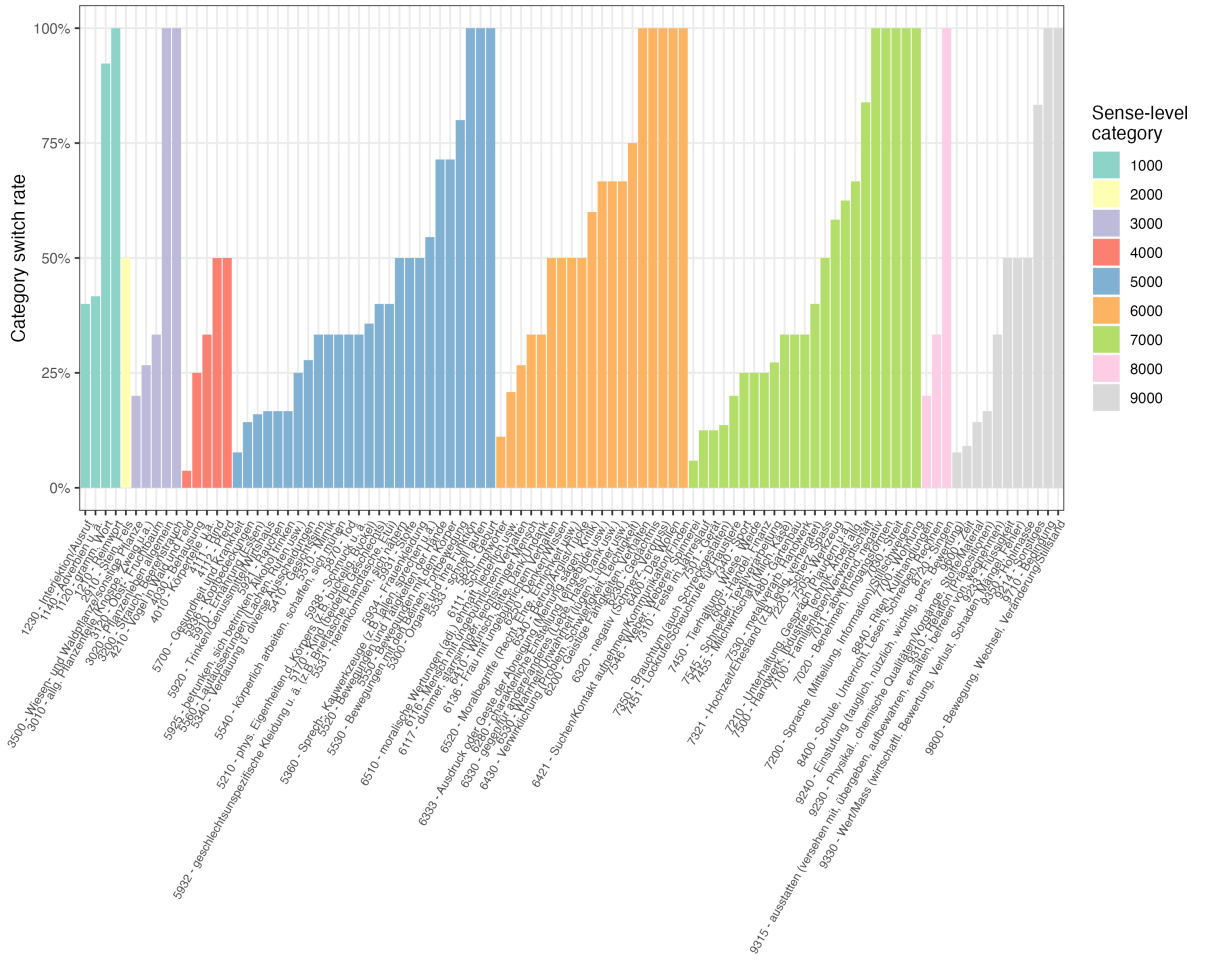

Figure 6: Switching rate of sense-level assignments between evaluation levels for the 1st and 2nd run; only sense-levels with a switching rate > 0 are shown. Bars represent individual sense-levels, while colors encode their higher-level semantic categories.

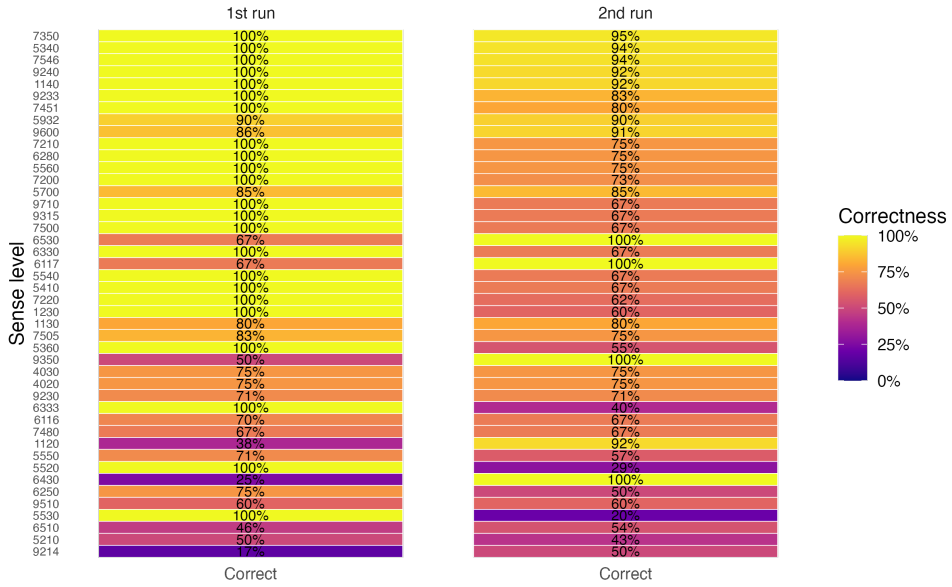

Figure 7: Correctness of sense-levels at the evaluation level *correct* across the 1st and the 2nd run; shown are only those sense-levels that exhibit volatility either within a run or between runs and that are evaluated as *correct* in at least one of the two runs. Sense-level labels are detailed in Figure 8.

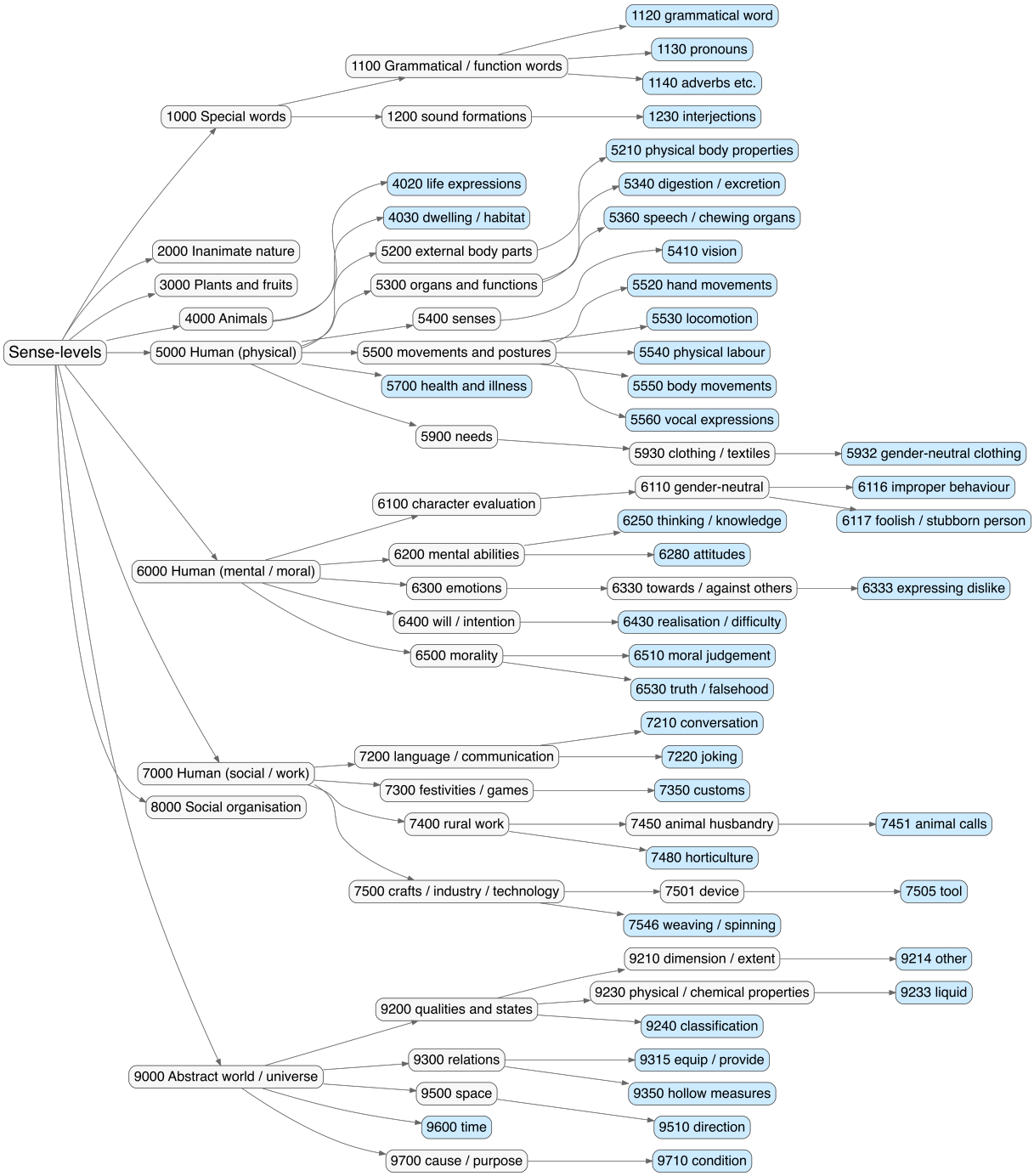

Figure 8: Taxonomic positions of sense-levels exhibiting volatility at the evaluation level *correct*. The same set of sense-levels as in Figure 7 is shown, positioned within the semantic taxonomy following Post (2026). Blue nodes mark the volatile endpoints of the taxonomic paths.

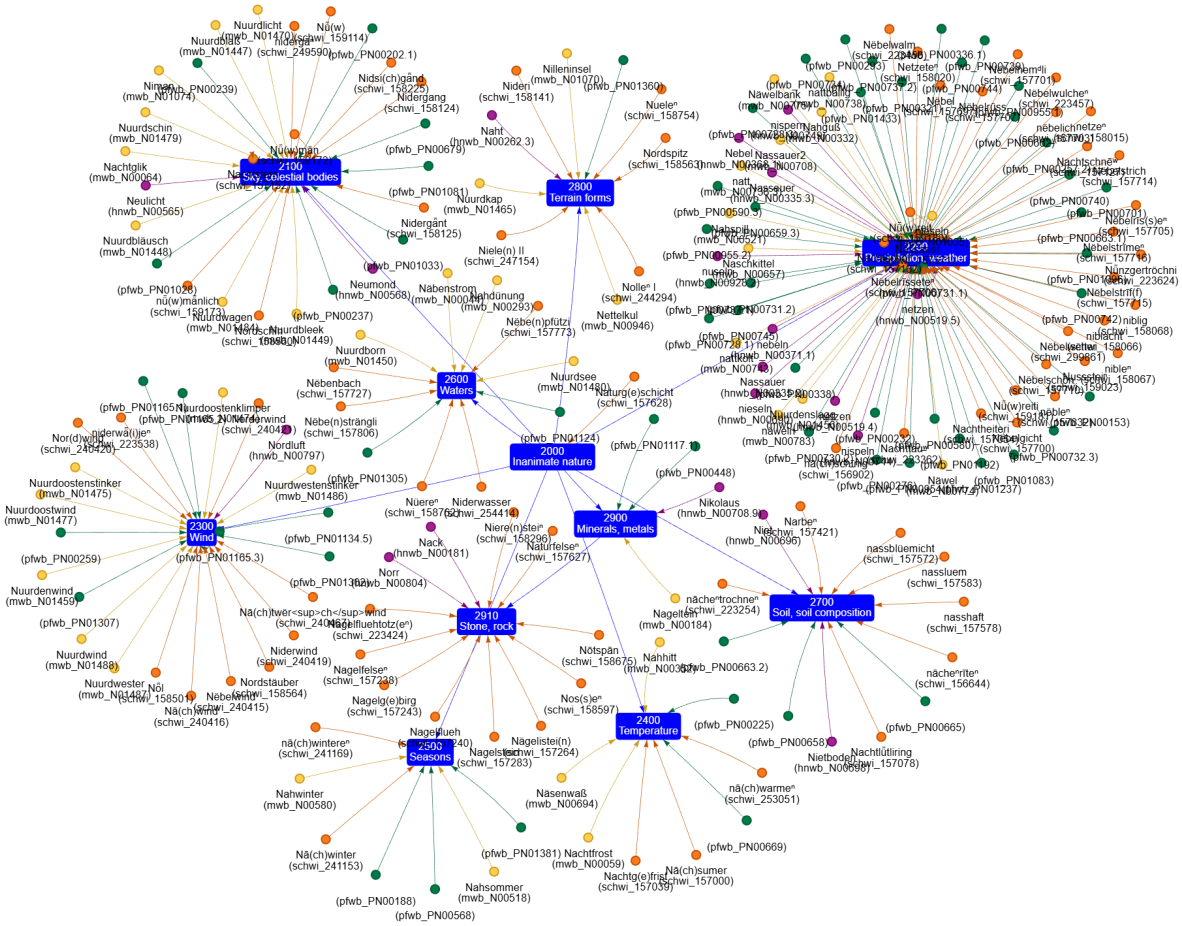

Figure 9: Semantic network based on the semantic taxonomy following (Post, 2026), showing clustering of sense-level entries across all subcategories under node 2000 “Inanimate Nature”, with assignments from the 1st, 2nd, and 3rd run merged into a shared conceptual space.

Role and Goal

- You are a linguistic classification model for historical lexicography.
- Task: Assign exactly one Hallig-Wartburg category (four-digit taxonomy ID) to each input line (an entry/sense with unique article_id).
- Think silently and return exclusively the required output format.

Context and Resources

- Binding taxonomy: Four-digit ID → Short designation. The taxonomy is authoritative and overrides external world knowledge. Use only IDs from this list; do not invent IDs.

[...]

Block 1 – Understanding and Preprocessing

- Parse the JSON fields. Use meaning (meaning specification) and translation_german (High German translation) as primary semantic source.
- Use lemma/lemma variants/grammatik information/ translation_latin(Latin, scientific designation) as secondary cues; draw morphological conclusions only if the lemma is unambiguously identified as Standard German. [...]

Block 2 – Feature Extraction → Candidates → Evaluation (incl. Consistency and Ambiguity)

[...]

- Derive suitable taxonomy candidates from the signals: first broad domain, then descend to the specifically matching leaf node (four-digit ID). Generate 3–5 candidates internally. Prefer more specific categories over coarser categories.

[...]

Block 3 – Decision and Output

- Select the best-matching specific leaf node (exactly one four-digit ID).
- Output exactly one output line for each entry in the same order. No use of conversation history.

Strict Output Format

- Format per Line: 'article_id':'taxonomie_id'
- taxonomie_id: exactly four digits (0–9), preserve leading zeros.
- No additional text, no explanations, no blank lines, no sorting, no "unknown".

[...]

Figure 10: Prompt extract used for sense-level alignment.