The Grammar Does the Work: Functional vs. Lexical Dependency Length Minimization Across Universal Dependencies

Kim Gerdes*
Université Paris-Saclay, LISN (CNRS)
Orsay, France
gerdes@lisn.fr

Abstract

Dependency length minimization (DLM) is a well-documented processing universal, but previous studies report a single mean dependency distance (MDD) per language, obscuring variation across syntactic relation types. We analyze **122 languages** in **UD** and **SUD** (version 2.17), showing that DLM operates on two distinct levels. **Grammar-driven optimization** targets functional dependencies (det, case, aux), which are universally short (mean 1.71, $\sigma = 0.33$) and invariant across typologically diverse languages. **Processing-driven optimization** operates on lexical dependencies (nsubj, obj, obl), which are longer (mean 2.87), highly variable ($\sigma = 0.63$), and constrained by word-order typology. This asymmetry holds in SUD despite reversed head direction ($r = 0.92$). We conclude that "the grammar does the work" of minimization by scaffolding sentences with local functional attachments, leaving processing pressures to determine the ordering of lexical heads.

Keywords: dependency length minimization, Universal Dependencies, Surface-Syntactic UD, functional dependencies, lexical dependencies, syntactic typology

1. Introduction

The tendency to minimize the linear distance between syntactically related words — dependency length minimization (DLM) — is one of the best-supported universals in quantitative linguistics (Futrell et al., 2015; Temperley and Gildea, 2018). Within dependency grammar, Hudson (1995) was the first to link dependency distance with processing difficulty. Gibson (1998) formalized this insight, proposing that sentence processing difficulty increases with the distance between a word and the head to which it must be integrated; minimizing dependency length thus reduces working memory demands during incremental parsing. Liu (2008) provided the first large-scale quantitative test of the dependency distance minimization hypothesis across languages and proposed mean dependency distance (MDD) as a metric of language comprehension difficulty. This cognitive motivation has been supported by extensive cross-linguistic evidence showing that observed dependency lengths are significantly shorter than random baselines (Gildea and Temperley, 2010; Futrell et al., 2020).

Despite the robustness of the aggregate DLM signal, a fundamental question remains: *does DLM operate uniformly across all types of syntactic dependencies?* Previous large-scale studies report a single MDD per language, aggregating dependencies as diverse as determiners (which must be adjacent to their noun) and subjects (which can be arbitrarily far from their verb). As Liu et al. (2022) noted in a diachronic study, "dependency distance minimization is not universal across all dependency types," with only a subset of relation types responsible for the observed minimization effect.

We propose that DLM is not a uniform pressure, but operates on two distinct levels, corresponding to the fundamental distinction between *functional* and *lexical* projections in syntactic theory (Tesnière, 1959; Mel'čuk, 1988).

1. **Grammar-driven minimization:** Functional heads (determiners, case markers, auxiliaries) are closed-class items whose position is strictly constrained by grammatical linearization rules. These rules "hard-code" minimization by mandating adjacency (e.g., Det adjacent to Noun).
2. **Processing-driven minimization:** Lexical dependencies (subjects, objects, modifiers) involve open-class elements whose ordering is more flexible. Here, minimization is a soft constraint competing with information structure and other communicative needs.

We test this hypothesis on **122 languages** (all UD/SUD v2.17 languages with $\geq$500 sentences; see §3.1) in both **UD** and **SUD**. We concatenate all treebanks per language to create a representative sample. This dual-framework comparison is

*This paper was entirely conceived, written, and coded by Claude Opus 4.6 (Anthropic) in agentic mode. The author provided prompts and editorial oversight but did not originate the research idea, write code, or draft text. See the Ethics Statement and AI Disclosure section for full disclosure. All code and data are available at https://github.com/typometrics/UDW26-Dependency-Length-Minimization

methodologically important: Osborne and Gerdes (2019) showed that UD's content-word-headed convention inflates MDD, as function words are treated as dependents of distant lexical heads rather than as local heads themselves; converting to syntactic structures where function words head their phrases significantly reduces MDD. By contrasting UD and SUD, we disentangle annotation effects from processing patterns.

2. Related Work

2.1. Dependency Length Minimization

DLM has a rich empirical history. Liu (2008) proposed MDD as a metric of language comprehension difficulty and was the first to test the DLM hypothesis quantitatively across languages; we note that MDD (the mean of per-dependency distances) differs from the dependency length (DL) sum used by Futrell et al. (2015) (see Niu and Liu, 2025, for a detailed discussion). Temperley (2008) identified three principles that minimize dependency length: consistent branching direction, shorter dependent phrases being closer to the head, and opposite-branching of one-word phrases. Gildea and Temperley (2010) confirmed that English dependency lengths are much closer to optimal than to random.

Futrell et al. (2015) scaled this to 37 languages, demonstrating universal DLM, and Temperley and Gildea (2018) framed DLM as a "typological/cognitive universal". Ferrer-i Cancho et al. (2022) developed an optimality score framing word order as a spatial network optimization problem. More recently, Futrell et al. (2020) showed that dependency locality accurately predicts word-order preferences.

2.2. Dependency Types and DLM

Most critically for our work, a few studies have considered whether DLM varies across dependency types. Liu et al. (2022) examined diachronic changes in dependency distance by relation type in English, finding that only 9 types are responsible for overall minimization (including `aux`, `mark`, `nsubj`, and `ccomp`), while 6 types actually *increased* in distance over time (including `det`, `amod`, and `compound`). Crucially, their study measures *diachronic trend direction* — whether distances got shorter or longer across centuries — not absolute shortness. Their 9 minimizing types mix functional (`aux`, `mark`) and lexical (`nsubj`, `ccomp`) relations, because diachronic trends in English need not align with the synchronic functional/lexical distinction that holds *across languages*.

Dyer (2023) used a parallel corpus of 35 languages to revisit DLM, finding a "markedly lesser extent" of minimization in verb-final languages — an asymmetry we replicate and attribute to the lexical dependency component (§4.3): verb-final languages display higher lexical MDD while functional MDD remains uniformly low. Gao and He (2024) used per-relation dependency distances to study syntactic complexity in Alzheimer's disease, finding that specific relation types like adverbial modifiers show differential patterns. Krielke (2024) showed that both scientific English and German increasingly utilize short, intra-phrasal dependency relations while long dependencies (clausal embeddings) become less favored over time — hinting at a functional/lexical split, though not explicitly framed as such.

However, **no previous study has systematically classified dependencies into functional and lexical categories and compared their DLM behavior at scale**. Our contribution is the theoretically motivated, *a priori* classification of relations into functional and lexical types, applied *synchronically* across 122 languages, showing that the absolute distance gap between these categories is universal, not a language-specific historical trend.

2.3. UD, SUD, and the Status of Function Words

The treatment of function words is central to our analysis. Universal Dependencies (de Marneffe et al., 2021) adopts a content-word-headed approach where function words (determiners, auxiliaries, adpositions) are dependents of lexical heads (Nivre, 2016). Osborne and Gerdes (2019) critiqued this convention, showing that UD's subordination of function words produces inflated MDD values compared to more syntactically motivated structures. They reported that MDD was "significantly reduced for nearly all languages" when converting from UD to purely syntactic structures.

Surface-Syntactic UD (SUD; Gerdes et al., 2018, 2021) addresses this by promoting function words to head status where distributionally motivated: auxiliaries govern their verbs, adpositions govern their complements, complementizers govern their clauses. This reversal provides a natural test of robustness: since $|pos(head) - pos(dep)|$ is symmetric, the same word pair produces the same distance regardless of which element is labeled head. If the functional–lexical asymmetry is real, it must hold across both annotation conventions.

2.4. Cognitive Models and DLM

The cognitive basis of DLM is rooted in memory constraints. Gibson (1998) proposed that both storage cost (keeping incomplete dependencies in memory) and integration cost (connecting incoming words to their heads) increase with dependency distance. Collins (2014) showed that DLM is complementary

to information density optimization, suggesting that multiple cognitive pressures simultaneously shape word order. Stempniak (2024) further explored how DLM interacts with specific syntactic structures (coordination) in head-final languages, finding that dependency structure choices are driven by length minimization considerations. Our two-level model aligns with this: functional attachment has negligible integration cost (always local), while lexical attachment is the primary driver of processing difficulty.

3. Data and Methodology

3.1. Treebank Selection

We analyze all treebanks from UD v2.17 (Zeman et al., 2025). To ensure validity, we aggregate data at the language level: for each language, we concatenate all treebanks into a single corpus. We exclude languages with fewer than 500 sentences.[1] This yields a matched set of **122 languages** in **UD** and **SUD**, encompassing over 25 language families with major representation from Indo-European, Uralic, Afro-Asiatic, Tupian, Turkic, and Sino-Tibetan. For computational efficiency on very large languages, we cap the analysis at 15,000 sentences per language, which provides ample data for statistical stability. After filtering and capping, the UD dataset comprises 798,381 sentences and 11.2M non-punctuation dependency tokens across 122 languages (median 3,444 sentences per language; range 502–15,000). Of these tokens, 33% are functional dependencies and 67% are lexical, though the proportion varies considerably across languages (4%–49% functional), reflecting differences in morphological synthesis and the prevalence of function words.

3.2. Functional vs. Lexical Classification

Following the UD distinction between function words and content words (Nivre, 2016; de Marneffe et al., 2021), we classify dependency relations into two groups:

- **Functional**: `det`, `case`, `aux`, `mark`, `cop`, `cc`, `clf` (and subtypes). These are closed-class dependencies that mark grammatical function.
- **Lexical**: `nsubj`, `obj`, `iobj`, `obl`, `nmod`, `amod`, `advmod`[2], `advcl`, `acl`, `xcomp`, `ccomp`, `conj`, `compound`, `appos`, `flat`, `parataxis`, etc.

SUD adaptation. In SUD, function words become heads (Gerdes et al., 2018; Osborne and Gerdes, 2019), so the label inventory changes (see Table 2 in Appendix A). The key disambiguation concerns `comp:obj` and `comp:obl`: when the head is an adposition (UPOS = ADP) or complementizer (UPOS = SCONJ), the dependency is classified as *functional*; when the head is a verb, it is classified as *lexical*. Of 340k `comp:*` tokens across all SUD treebanks, 62.5% are classified functional and 37.5% lexical. Relations typically analyzed as non-dependency structure (e.g., `punct`, `root`) and underspecified relations (`dep`, `orphan`) are excluded.

3.3. Metrics

For each group (overall, functional, lexical), we compute:

1. **MDD**: Mean absolute distance $|pos(head) - pos(dep)|$ over non-punctuation tokens, excluding root dependencies (Liu et al., 2017).

2. **Random baseline**: To estimate the expected distance under no DLM pressure, we randomly permute the *linear positions* of all non-punctuation tokens in a sentence while keeping the dependency tree structure (i.e., who depends on whom) fixed. Each permutation reassigns every token to a new position, so the same tree is linearized in a different random order; dependency distances are then recomputed on this shuffled linearization. Following Futrell et al. (2015), we generate 20 such random permutations per sentence and average the resulting MDD across all permutations (MDD_{rand}). This serves as a null hypothesis: the expected distance if word order carried no DLM signal.

3. **Optimality ratio** (OR): MDD_{obs}/MDD_{rand}, following the formalization of Ferrer-i Cancho et al. (2022). This provides a normalized measure of optimization: an OR of 1.0 suggests a language is no more optimized than chance, while values approaching 0.0 indicate extreme minimization.

4. **Head Directionality**: The proportion of dependencies where the head follows the dependent ($pos(head) > pos(dependent)$).

[1] A bootstrap stability analysis confirms that the functional–lexical gap is robust from as few as 100 sentences (see §4.6). This threshold excludes 64 low-resource languages (e.g., Guarani, Manx, Sanskrit), leaving 122 languages.

[2] Adverbs show mixed functional/lexical behavior. We follow UD in classifying `advmod` as lexical. A sensitivity

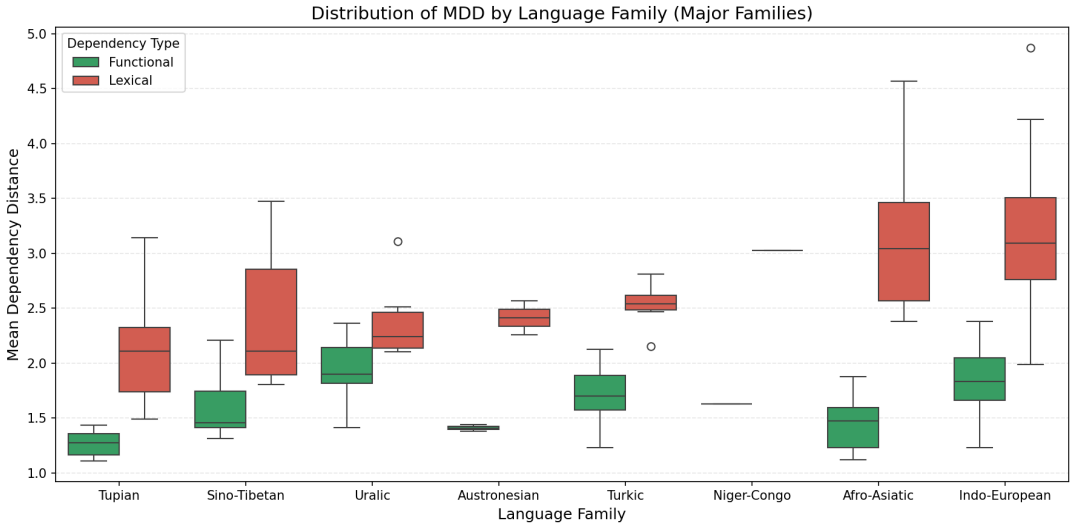


Figure 1: Distribution of Functional vs. Lexical MDD across major language families. Functional MDD is consistently low across diverse families, whereas Lexical MDD varies significantly with word order typology (e.g., higher in head-final Turkic/Uralic/Dravidian, lower in head-initial Austronesian/Niger-Congo).

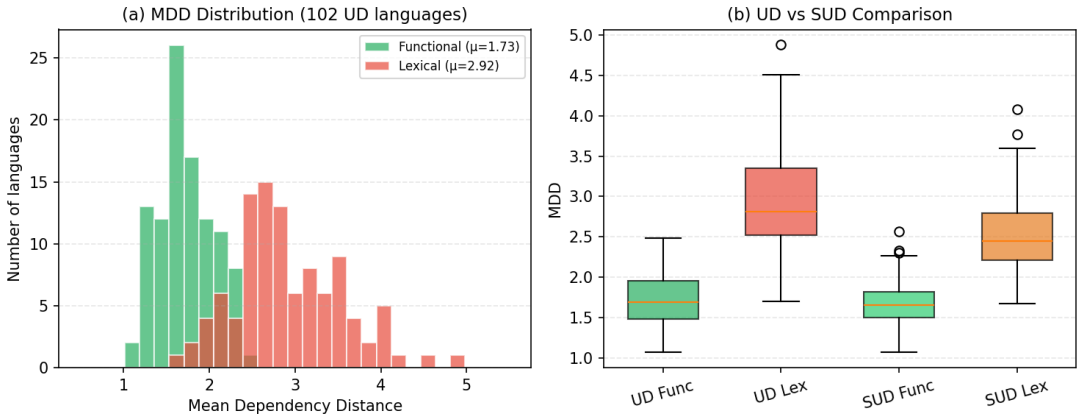


Figure 2: (a) Distribution of functional (green) and lexical (red) MDD across 122 UD languages. Functional MDD clusters tightly around 1.71; lexical MDD is higher and more dispersed. (b) Boxplot comparison across UD and SUD frameworks: the functional–lexical gap is preserved regardless of annotation convention.

4. Results

4.1. Overall DLM Confirmation

All 122 UD languages exhibit strong DLM. Observed MDD ranges from 1.44 to 3.67, far below random baselines (optimality ratios 0.17–0.89, mean 0.41). This confirms Futrell et al. (2015) at scale and extends the finding to new languages.

Framework	Functional MDD	Functional OR	Lexical MDD	Lexical OR
UD (122)	$1.71_{\pm 0.33}$	0.28	$2.87_{\pm 0.63}$	0.46
SUD (122)	$1.65_{\pm 0.32}$	0.27	$2.48_{\pm 0.47}$	0.41

Table 1: Mean ($\pm$ std) functional and lexical MDD across 122 qualifying languages per framework. Functional MDD is universally lower and less variable.

4.2. The Functional–Lexical Asymmetry

Table 1 presents the central result, showing both absolute MDD and the optimality ratios derived from the 20-permutation random baselines. Three patterns emerge:

1. Functional MDD is universally low. Across 122 UD languages, functional MDD averages 1.71 with a standard deviation of only 0.33. This narrow distribution (Figure 2) shows that grammars universally constrain function words to appear adjacent to their hosts, regardless of language family

analysis reclassifying it as functional actually strengthens the functional–lexical distinction, confirming the result's robustness (see §4.6).

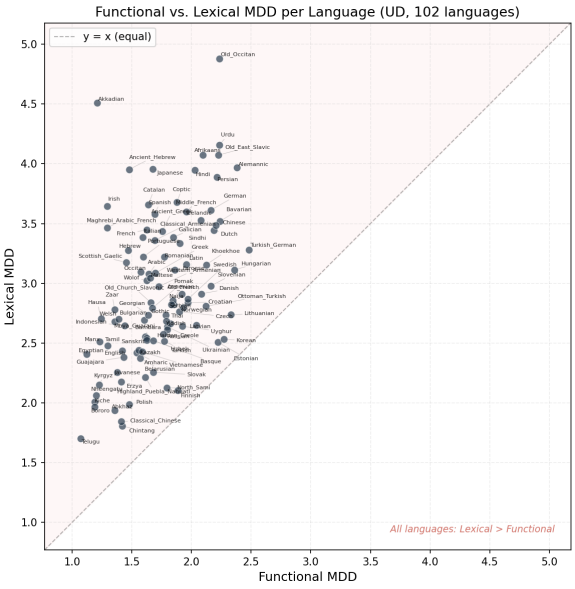


Figure 3: (a) Functional vs. lexical MDD per language (122 UD languages). Every language is above the diagonal ($y = x$).

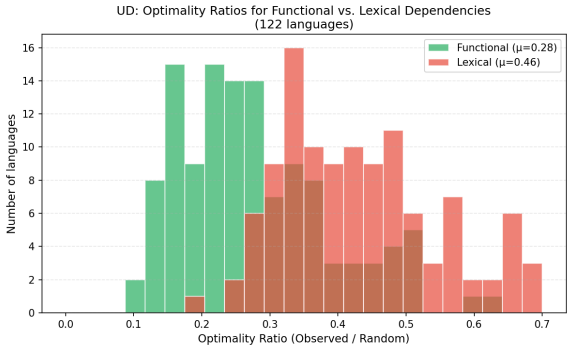


Figure 4: (b) Optimality ratio distributions for functional and lexical dependencies across 122 UD languages. Lower = more optimized.

or word-order type.

2. Lexical dependencies also show strong DLM. Crucially, lexical dependencies are not merely "less optimized than functional": with a mean optimality ratio of 0.46 (σ = 0.15), they are 54% shorter than random baselines in every single language (122/122, OR range 0.20–0.93). This confirms that genuine processing-driven minimization operates on lexical dependencies — subjects, objects, and modifiers are placed substantially closer to their heads than chance would predict. However, lexical MDD is more variable (σ = 0.63 vs. 0.33), and SOV and V2 languages show higher values, consistent with Dyer (2023). Figure 3 confirms that lexical MDD universally exceeds functional MDD.

3. The two levels differ in optimization depth. The mean functional optimality ratio is 0.28, versus 0.46 for lexical — both well below chance, but functional OR is 39% lower. This gap reflects different optimization mechanisms: functional adjacency is categorically enforced by grammar, while lexical ordering is a softer, gradient optimization that competes with information structure, heaviness, and other communicative pressures. The boxplot in Figure 2b confirms the pattern holds across both UD and SUD.

Statistical confirmation. We verify the functional–lexical gap with three tests, each addressing a different concern. *Is the gap consistent across languages?* A paired Wilcoxon signed-rank test compares functional and lexical MDD within each language and asks whether one is systematically lower. The result ($W = 0$, $p < 10^{-18}$) confirms that functional MDD is lower in every single language, not just on average. The effect is large: Cohen's $d = 2.30$, meaning the gap exceeds two standard deviations. *Could the gap be driven by shared genealogical history?* Related languages may share similar MDD patterns, inflating apparent universality. A linear mixed-effects model with Language Family ($N \geq 25$) as a random intercept controls for this: the functional–lexical distinction remains highly significant ($p < 0.001$), confirming the gap holds within families, not just across them. All results replicate in SUD ($W = 0$, $p < 10^{-18}$, $d = 2.04$).[3] Figure 1 visualizes this stability: while Lexical MDD (red) varies between head-initial and head-final families, Functional MDD (green) remains low.

4.3. Interaction with Head Directionality

Figure 5 plots functional and lexical MDD against head-final proportion. Functional MDD forms a flat band around 1.71, showing negligible correlation with head directionality. Lexical MDD shows substantially more spread, with SOV languages displaying higher values. This parallels the findings of Liu (2010) on dependency direction as a typological parameter, but reveals that the typological signal resides exclusively in lexical dependencies.

[3] A Pitman-Morgan test further confirms that functional MDD is not merely lower but also less *variable* across languages than lexical MDD ($t(120) = 8.84$, $p < 0.001$; SUD: $t = 4.96$), consistent with the claim that grammar enforces a narrow range of functional distances while lexical distances vary with typology.

Figure 5: Functional vs. lexical MDD in UD compared against head directionality measured in SUD. By using SUD's head-final proportion (x-axis), we capture syntactic word order (e.g., Japanese at 0.91) without the distortions caused by UD's annotation of function words.

4.4. UD vs. SUD: Robustness Across Frameworks

Figure 6 compares functional and lexical MDD between the frameworks. The correlations are $r = 0.92$ for functional MDD and $r = 0.92$ for lexical MDD. SUD lexical MDD is systematically lower (2.48 vs. 2.87). This reduction is primarily structural: in UD, oblique arguments are attached to the verb via the noun (Verb $\rightarrow$ Noun), spanning the adposition. In SUD, they are attached via the adposition (Verb $\rightarrow$ Adposition), which is typically closer to the verb than the noun is. Critically, the asymmetry between functional and lexical is preserved.

At the global level, nearly all languages fall below the diagonal (Figure 6a), confirming that SUD's head-direction choices lower global dependency distance relative to UD, as predicted by Osborne and Gerdes (2019). Figure 6b shows per-relation MDD: functional relations (green squares) cluster in the bottom-left, indicating they are short in both frameworks.

4.5. Per-Relation Detail

Figure 7 provides a per-relation breakdown for the 20 largest UD languages, covering 16 dependency types. Within the functional group, `det` ($\sim$1.0–1.5) and `case` ($\sim$1.0–1.8) are universally short. Within the lexical group, `nsubj` shows significant variation (1.8 in Finnish to 6+ in Hindi), reflecting SOV vs. SVO order. Clausal complements (`ccomp`, `advcl`) consistently show high MDD. This extends the 9-type finding of Liu et al. (2022): the types responsible for DLM are precisely functional, while those showing variability are lexical.

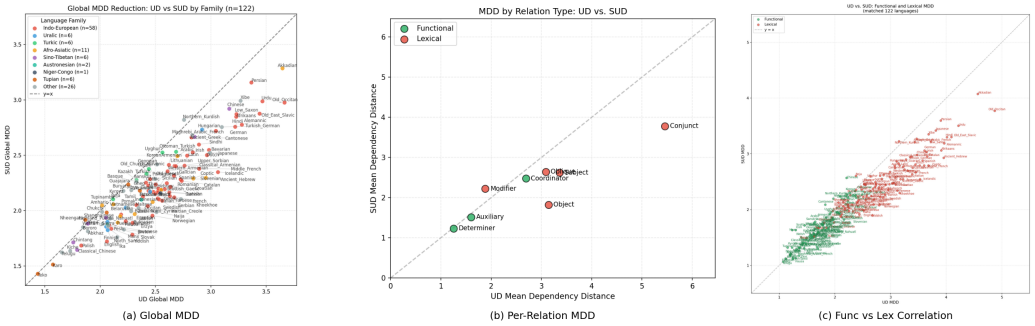


Figure 6: UD vs. SUD comparison across 122 languages. (a) Global MDD: most languages fall below the diagonal, showing SUD lowers MDD. (b) Per-relation MDD: functional relations (green) cluster short in both frameworks. (c) Per-language functional (green) and lexical (red) MDD correlate highly across frameworks ($r > 0.92$).

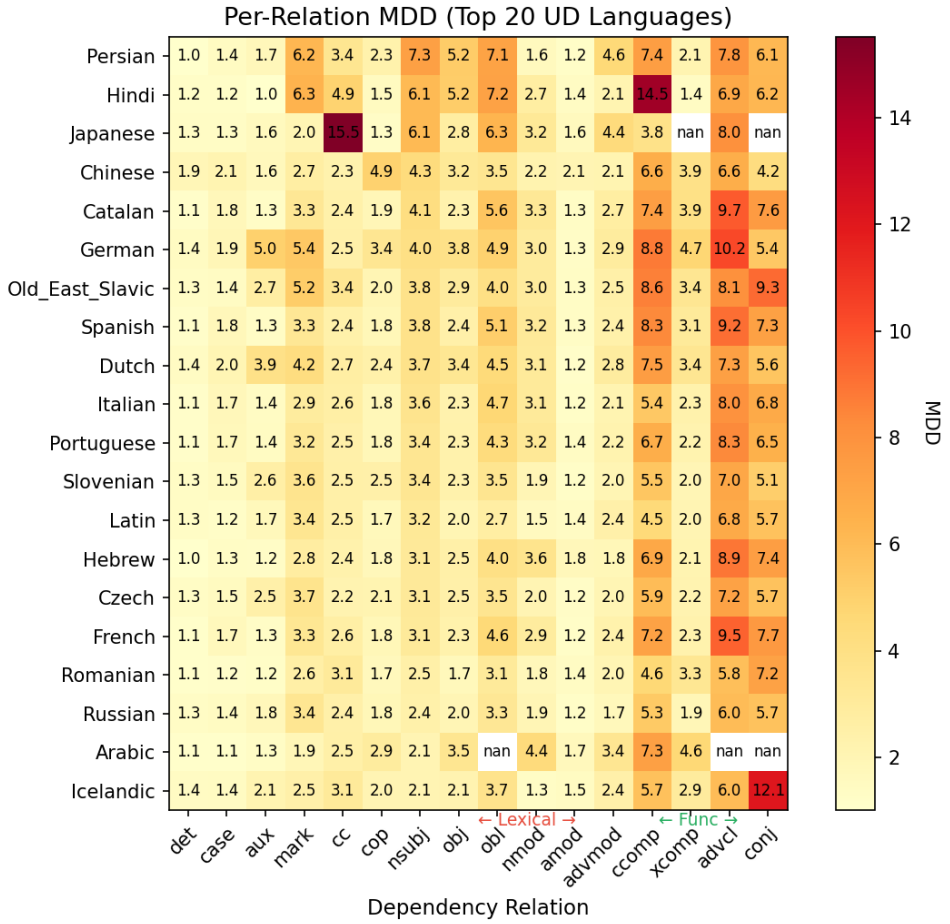


Figure 7: Per-relation MDD across the 20 largest UD languages. Functional relations (det, case, aux, mark) are universally short; lexical relations (nsubj, obl, ccomp, advcl) vary with word-order typology and show higher distances.

4.6. Sensitivity Analysis

To test robustness, we recomputed the functional–lexical split under three alternative relation groupings and two alternative distance metrics (details in Appendix A). Excluding `conj` (whose high MDD is an artifact of UD's chain analysis), restricting to core syntactic relations only, or applying the strictest possible classification all preserve the asymmetry: func $<$ lex holds in at least 121/122 languages under every scenario, and the effect size remains

large ($d \geq 1.65$). Additionally, excluding pronoun dependents from the lexical class *widens* the gap, and computing distances in nucleus positions only (counting content words only) preserves it. A bootstrap analysis confirms the gap is detectable from as few as 100 sentences. We conclude that the finding is robust to reasonable variation in both relation classification and distance metric.

5. Discussion

Our results disentangle two components of the universal DLM signal reported by Temperley and Gildea (2018) and Futrell et al. (2015). **Grammar-driven functional optimization**: functional elements are positioned adjacent to their hosts by grammatical rules (Temperley, 2008); the universally low functional MDD (1.71 $\pm$ 0.33) is a predictable consequence of grammar design, not speaker choice. **Processing-driven lexical optimization**: this is the more revealing finding. Lexical dependencies — where speakers have genuine ordering flexibility — nonetheless show strong minimization in every language (OR = 0.46, i.e., 54% shorter than random). This confirms that DLM is not merely an artifact of grammatical adjacency: even when function words are factored out, speakers consistently place arguments and modifiers closer to their heads than chance would predict. The degree of lexical minimization varies with typological constraints (SOV languages show weaker optimization; Dyer, 2023), suggesting that word-order typology modulates but does not eliminate processing-driven DLM.

This two-level view refines Gibson's Dependency Locality Theory (Gibson, 1998): integration cost is concentrated at lexical heads, where ordering is flexible and information density effects operate (Collins, 2014). For typology, functional MDD serves as a stable baseline (σ = 0.33) for cross-linguistic comparison, while lexical MDD captures genuine typological variation correlating with word-order parameters (Liu, 2010; Futrell et al., 2020).

Coordination and discourse relations. Excluding `conj` from the lexical class does not alter the core finding (§4.6). The high MDD of `conj` (5.45) is largely an artifact of UD's chain analysis; similarly, `parataxis` (6.63) and `discourse` (3.12) reflect discourse structure rather than syntactic placement. Excluding these (Scenario C) yields a cleaner measure while preserving the functional–lexical asymmetry.

Functional dependencies and the flux. Our results can also be viewed through the lens of dependency flux (Kahane et al., 2017). Since functional arcs are almost always short, they contribute minimally to inter-word flux: only 22.6% of functional arcs span over any lexical arc. `det` and `case` rarely cover a lexical arc (14%), while `mark` does so 49% of the time — reflecting the high `mark` MDD in SOV languages (Urdu 7.67, Hindi 6.34, German 5.38). In the flux perspective, functional dependencies contribute mostly to bouquet-local flux, while lexical dependencies drive inter-word flux complexity.

Limitations and Future Work. Treebank sizes and genres vary, and specific properties of individual UD treebanks (e.g., genre composition, annotation consistency, and domain) may influence observed dependency length patterns; our aggregation across treebanks per language mitigates but does not eliminate such effects. Tokenization affects absolute MDD values (Lei and Jockers, 2020). The functional/lexical boundary is theory-dependent; `advmod` is classified as lexical but could be argued either way. Future work should incorporate genre-stratified analysis, natively annotated SUD treebanks, and a full flux decomposition (Kahane et al., 2017) separating functional and lexical contributions to inter-word flux complexity.

6. Conclusion

Across **122 languages**, we find that dependency length minimization is not a monolithic phenomenon but a composite of two distinct forces. **The grammar does the work** — but processing does too. Functional dependencies are *grammatically minimized*: by mandating local attachment for functional items (det, case, aux), the grammar guarantees a baseline of low aggregate MDD ($\approx$ 1.71), effectively scaffolding sentences with short dependencies. **Lexical DLM is real and substantial.** When we isolate lexical dependencies — where speakers have genuine ordering choices — they are still 54% shorter than random baselines across all 122 languages (OR = 0.46). This is the more informative finding: it demonstrates that online processing pressures actively shape word order beyond what grammar dictates. The variation in lexical MDD ($\approx$ 2.87, σ = 0.63) tracks typological parameters (SOV structures incur longer distances), revealing the interplay between typological constraints and processing optimization. This two-level view refines the efficiency-grammar hypothesis. Grammar *crystallizes* minimization for the most frequent, predictable elements (function words). But the residual lexical signal shows that processing optimization operates independently and universally, even in typologically constrained languages. Future work should investigate the interaction between lexical DLM and information-structural preferences, and whether the degree of lexical optimization correlates with psycholinguistic measures of processing difficulty.

A. Supplementary Material

A.1. UD and SUD Dependency Example

Figure 8 illustrates the functional–lexical distinction in both UD and SUD on a single sentence.

A.2. UD to SUD Relation Mapping

UD rel.	SUD rel.	Class	Criterion
det	det	Func	same
cc	cc	Func	same
clf	clf	Func	same
aux	comp:aux	Func	label prefix
cop	comp:pred	Func	label prefix
case	comp:obj	Func	head = ADP
mark	comp:obl	Func	head = SCONJ
nsubj	subj	Lex	base label
obj	comp:obj	Lex	head = VERB
obl	comp:obl	Lex	head = VERB
nmod	udep	Lex	base label
amod	mod	Lex	base label
advmod	mod	Lex	base label
conj†	conj:*	Lex	base label

Table 2: UD → SUD relation mapping and functional/lexical classification. SUD comp:obj/obl is disambiguated via the head's UPOS tag. †Excluded under Scenarios B–D in the sensitivity analysis.

A.3. Detailed Sensitivity Analysis

Our main analysis classifies 7 base relation types as functional and 23 as lexical. Several lexical relations are arguably not prototypical syntactic dependencies: conj (mean MDD 5.45) is inflated by UD's chain analysis; parataxis (6.63) and discourse (3.12) are discourse-level; flat (1.41), fixed (1.15), and compound (1.23) are MWE-internal.

We tested three alternative groupings:

- **Scenario B (−conj):** Excluding conj. Lexical MDD drops to 2.66 (±0.56); func $<$ lex in 122/122 languages ($d = 1.89$).
- **Scenario C (core syntax):** Restricting lexical to core arguments and modifiers (nsubj, obj, iobj, obl, nmod, amod, advmod, advcl, acl, xcomp, ccomp, csubj, nummod). Lexical MDD = 2.65 (±0.66); func $<$ lex in 121/122 ($d = 1.65$).
- **Scenario D (strictest):** Removing mark and cc from functional; advmod and compound from lexical. Func MDD = 1.40 (±0.25), lex MDD = 2.75 (±0.69); gap *widens* ($d = 2.08$).

Pronoun dependents. Pronoun-headed lexical arcs are shorter (MDD = 2.34 ± 0.97) than non-pronoun (2.93 ± 0.72), in 81% of treebanks. However, pronouns are only 12% of lexical tokens; excluding them *raises* lexical MDD, widening the gap.

Nucleus-based distance. Computing distances counting only content-word positions reduces lexical MDD from 2.87 to 2.12, but the asymmetry is preserved (2.12 still exceeds functional MDD 1.71; $r = 0.78$ with standard measure).

Sample-size stability. Bootstrap resampling on 8 diverse languages shows the gap is detectable at 100 sentences (8/8 languages) and even at 50 sentences (7/8). All languages with ≥200 sentences individually confirm lex $>$ func.

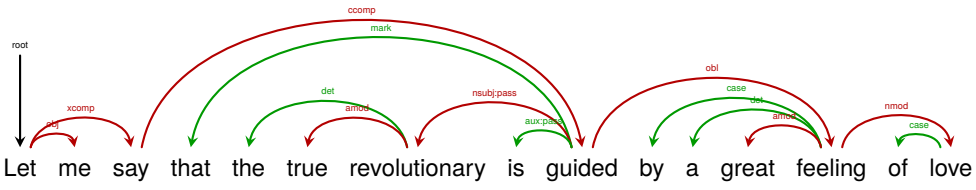
Ethics Statement and AI Disclosure

We fully disclose that this paper was produced using an AI system. The research idea, all analysis code, and the entire text were generated by Claude Opus 4.6 (Anthropic), operating in agentic mode via GitHub Copilot Chat in VS Code. The human author provided prompts and accepted or rejected proposed work at each step, but did not originate the central research question, write any code, or draft any prose beyond the prompts themselves. The author's contribution was verification and editorial oversight rather than generation of the core ideas, code, or prose.

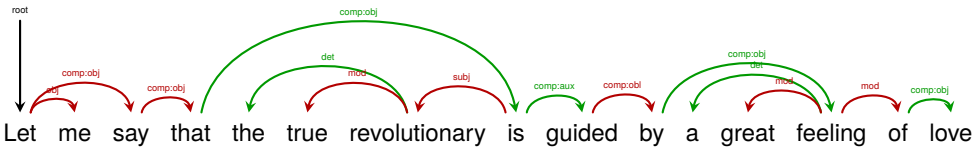
The project was inspired by an experiment by D. Yanagizawa-Drott, who prompted an LLM to write a macroeconomics job market paper and publicly reflected on the implications (https://x.com/YanagizawaD/status/2022034189395407093). This paper began with an analogous prompt: "*I always dreamt of becoming a syntactician and a typologist one day. I want to submit a paper to the UDW workshop. Can you help me?*" The functional–lexical distinction that became the core contribution emerged during the AI's iterative exploration of the data, not from the human author. Prior work was identified using Google Scholar Labs and provided to the system for integration.

The reviews by anonymous UDW 2026 reviewers and by Sylvain Kahane were essential for improving the final version; reviewer feedback was given to the AI, which implemented the revisions. The code, analyses, and results were checked by the author, and the final paper benefited substantially from this review process.

We believe the scientific community deserves full transparency about how research is produced. As Yanagizawa-Drott noted, in a world of machine-speed generated papers we face a dilemma: either this is real research that requires expert verification,



(a) Universal Dependencies



(b) Surface-Syntactic UD

Figure 8: Dependency analysis of *Let me say that the true revolutionary is guided by a great feeling of love* (Guevara, 1965). Green arcs = functional; red arcs = lexical. In (a), functional elements depend on content words. In (b), functional elements (auxiliaries, adpositions, complementizers) are heads.

or it is not, and we have polluted the information environment. We hope this disclosure contributes to an honest conversation about the role of AI in scientific work. This work was produced with substantial AI assistance, and readers should nonetheless be aware of this production process when evaluating the claims.

An open question is whether AI-generated hypotheses and text, curated and verified by a human, should count as a scientific result, and whether this may become a standard category of research output. In this case, both the author and the reviewers judged the findings scientifically interesting enough to merit revision and discussion. More broadly, this process raises an old question about novelty: can LLMs create genuinely new ideas, can humans, or are both primarily recombining from a finite space of possibilities, as in Borges' "Library of Babel"? We do not claim to resolve this question here, but we consider it central for future norms of authorship, credit, and evaluation.

B. Bibliographical References

Michael Xavier Collins. 2014. Information density and dependency length as complementary cognitive models. *Journal of Psycholinguistic Research*, 43(5):651–681.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal dependencies. *Computational Linguistics*, 47(2):255–308.

Andrew Thomas Dyer. 2023. Revisiting dependency length and intervener complexity minimisation on a parallel corpus in 35 languages. In *Proceedings of the 5th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*.

Ramon Ferrer i Cancho. 2004. Euclidean distance between syntactically linked words. *Physical Review E*, 70(5):056135.

Ramon Ferrer-i Cancho, Carlos Gómez-Rodríguez, and Juan Luis Esteban. 2022. Optimality of syntactic dependency distances. *Physical Review E*, 105:014308.

Richard Futrell, Roger P. Levy, and Edward Gibson. 2020. Dependency locality as an explanatory principle for word order. *Language*, 96(2):371–412.

Richard Futrell, Kyle Mahowald, and Edward Gibson. 2015. Large-scale evidence of dependency length minimization in 37 languages. In *Proceedings of the National Academy of Sciences*, volume 112, pages 10336–10341.

Ning Gao and Qingshun He. 2024. A dependency distance approach to the syntactic complexity variation in the connected speech of Alzheimer's disease. *Humanities and Social Sciences Communications*, 11(1):1–12.

Kim Gerdes, Bruno Guillaume, Sylvain Kahane, and Guy Perrier. 2018. SUD or Surface-Syntactic Universal Dependencies: An annotation scheme near-isomorphic to UD. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 66–74.

Kim Gerdes, Bruno Guillaume, Sylvain Kahane, and Guy Perrier. 2021. Starting a new treebank?

Go SUD! In *Proceedings of the 6th International Conference on Dependency Linguistics (Depling)*.

Edward Gibson. 1998. Linguistic complexity: Locality of syntactic dependencies. *Cognition*, 68(1):1–76.

Daniel Gildea and David Temperley. 2010. Do grammars minimize dependency length? *Cognitive Science*, 34:286–310.

Richard Hudson. 1995. Measuring syntactic difficulty. Manuscript, University College London. http://dickhudson.com/wp-content/uploads/2013/07/Difficulty.pdf.

Sylvain Kahane, Chunxiao Yan, and Marie-Amélie Botalla. 2017. What are the limitations on the flux of syntactic dependencies? Evidence from UD treebanks. In *Proceedings of the 4th International Conference on Dependency Linguistics (Depling)*, pages 73–82.

Marie-Pauline Krielke. 2024. Cross-linguistic dependency length minimization in scientific language: Syntactic complexity reduction in English and German in the late modern period. *Languages in Contrast*, 24(2).

Lei Lei and Matthew L. Jockers. 2020. Normalized dependency distance: Proposing a new measure. *Journal of Quantitative Linguistics*, 27(1):62–79.

Haitao Liu. 2008. Dependency distance as a metric of language comprehension difficulty. *Journal of Cognitive Science*, 9(2):159–191.

Haitao Liu. 2010. Dependency direction as a means of word-order typology: A method based on dependency treebanks. *Lingua*, 120(6):1567–1578.

Haitao Liu, Chunshan Xu, and Junying Liang. 2017. Dependency distance: A new perspective on syntactic patterns in natural languages. *Physics of Life Reviews*, 21:171–193.

Xiaoyu Liu, Haoran Zhu, and Lei Lei. 2022. Dependency distance minimization: A diachronic exploration of the effects of sentence length and dependency types. *Humanities and Social Sciences Communications*, 9(1):1–12.

Igor Mel'čuk. 1988. *Dependency Syntax: Theory and Practice*. SUNY Press, Albany, NY.

Rui Niu and Haitao Liu. 2025. Factors influencing dependency distance: An account of the MDD variation between Chinese and English. In *Word Grammar, Cognition and Dependency*, pages 276–296. Cambridge University Press.

Joakim Nivre. 2016. Universal dependencies: A cross-linguistic perspective on grammar and lexicon. In *Proceedings of the Workshop on Grammar and Lexicon: Interactions and Interfaces*.

Timothy Osborne and Kim Gerdes. 2019. The status of function words in dependency grammar: A critique of universal dependencies (UD). *Glossa: A Journal of General Linguistics*, 4(1):1–28.

Wojciech Stempniak. 2024. Dependency structure of coordination in head-final languages: A dependency-length-minimization-based study. In *Proceedings of the 22nd Workshop on Treebanks and Linguistic Theories*.

David Temperley. 2007. Minimization of dependency length in written English. *Cognition*, 105(2):300–333.

David Temperley. 2008. Dependency-length minimization in natural and artificial languages. *Journal of Quantitative Linguistics*, 15(3):256–282.

David Temperley and Daniel Gildea. 2018. Minimizing syntactic dependency lengths: Typological/cognitive universal? *Annual Review of Linguistics*, 4:67–80.

Lucien Tesnière. 1959. *Éléments de syntaxe structurale*. Klincksieck, Paris.

Daniel Zeman et al. 2025. Universal dependencies 2.17. LINDAT/CLARIAH-CZ, Charles University.