Case–Number Dissociation in Finnish Noun Embeddings: `fastText` vs. `BERT` Layer Effects

Alexandre Nikolaev
University of Eastern Finland
Joensuu, Finland
alexandre.nikolaev@uef.fi

Yu-Ying Chuang
National Taiwan
Normal University
Taipei, Taiwan
yuying.chuang@ntnu.edu.tw

R. Harald Baayen
University of Tübingen
Tübingen, Germany
harald.baayen@uni-tuebingen.de

Abstract

Motivated by how inflectional morphology is encoded in modern embeddings, we revisit the 55,271 inflected forms from the 2,000 most frequent Finnish nouns analyzed by Nikolaev et al. (2022) using fastText and ask a single question: *where does inflectional morphology emerge in* BERT*?* For each form, we extract *minimal-context* FinBERT vectors from every layer (1–12) by running each word in isolation and averaging its *WordPiece vectors* into a single representation. Using the same generating model as in Nikolaev et al. (2022), we impute latent vectors for the stem, NUMBER, CASE, POSSESSIVE, and CLITIC, plus a higher-order interaction, and evaluate by rank-1 nearest correlation.

Within BERT, accuracy follows an *emergence curve* from 67.21% (layer 1) to 86.16% (layer 12). The error mix shifts with depth: middle layers show a lower share of CASE errors but a higher share of NUMBER errors, whereas the top layer reverses this tendency; clitic-only errors are rare throughout. For context, the fastText ceiling is slightly higher (≈89%), but our focus is the layer-resolved profile inside BERT.

The result is a compact, reproducible map of Finnish noun inflection across the BERT stack, showing how different inflectional cues become recoverable at different depths (BERT layers) under an identical modeling and evaluation pipeline.

1 Introduction

We take the same 55,271 inflected forms derived from the 2,000 most frequent Finnish nouns in Nikolaev et al. (2022) and ask a single question: *where does inflectional morphology emerge in* BERT*?* Whereas Nikolaev et al. (2022) evaluated fastText (Bojanowski et al., 2017), we keep the items and pipeline unchanged but replace the target space with BERT, treating each BERT layer as a separate target space.

Nikolaev et al. (2022) introduced the simple idea we use here: treat each inflected form as a sum of a few “building blocks”, one vector for the stem (lexeme) and one vector for each inflectional feature (number, case, possessive, clitic), plus optional interaction blocks when features combine. Formally, a design matrix L says which blocks are “on” for each form; S holds the gold vectors; and we learn the block vectors Q by solving the linear system $LQ = S$ (least squares). A predicted form is then $\hat{S} = LQ$, and we score it by checking whether its nearest neighbour by correlation is the correct gold vector (“rank-1” accuracy).

Using fastText, Nikolaev et al. (2022) showed three key facts: adding *case* gives the first big jump in accuracy; a *number*×*case* interaction is required to capture non-additive structure; and adding the higher-order bundle (number:case:possessive:clitic) yields the best overall performance (about 89–92%). We keep *the same* items, the same design L, and the same evaluation, and ask how accuracy, the composition of errors, and the geometry of the space change *across* BERT *layers* under this identical setup.

Applied layer by layer (each BERT layer as its own target space), this reused model gives three concise diagnostics of “emergence”: (i) overall recoverability (accuracy of $\hat{S}^{(\ell)}$ across layers); (ii) combination sensitivity (gains from interaction terms at each layer); and (iii) feature fragility (within-layer error composition by category). Together these yield a layer-resolved map of inflectional morphology in BERT that is directly comparable to the fastText baseline.

For BERT, we extracted *minimal-context* vectors from the cased Finnish encoder (Virtanen et al., 2019; Devlin et al., 2019). For each surface form, we tokenized it with the FinBERT WordPiece tokenizer and constructed the minimal input `[CLS]` $t_1 \dots t_k$ `[SEP]`, where t_i are WordPiece segments (Schuster and Nakajima, 2012) (no additional con-

Proceedings of the 10th International Workshop on Computational Linguistics for Uralic Languages, pages 127–130
December 10-12, 2025 ©2025 Association for Computational Linguistics

text). We then ran a forward pass through the pretrained encoder with parameters held fixed (evaluation mode; dropout disabled; no fine-tuning) to obtain layer-wise hidden states, selected a layer $\ell \in \{1, \ldots, 12\}$, and mean-pooled the layer-ℓ vectors over the WordPiece positions, excluding `[CLS]` and `[SEP]`. This yielded one 768-dimensional vector per form *per layer*. We did *not* average across sentence occurrences. By contrast, a `fastText` type vector was a single parameter learned from all occurrences of a form and, via character n-grams, effectively summarized corpus-wide usage in one vector. Our BERT vectors are *usage-trained* in the sense that the encoder's parameters were learned from large Finnish text corpora using self-supervised objectives (masked-language modeling), so they encode distributional regularities of how forms occur across contexts. At extraction time, however, we supplied no surrounding words (only `[CLS] wordpieces [SEP]`) and mean-pooled a chosen layer over the wordpieces. The resulting vectors are deterministic, type-like summaries that reflect the model's usage-trained knowledge without being conditioned on any specific sentence. We adopted this minimal-context setting to localize *where* inflectional cues resided across layers while holding items and evaluation fixed. Averaging BERT token vectors over many sentences would have made them more `fastText`-like as type proxies, but it would have introduced corpus/sense sampling choices and mixed context effects with layer effects; we therefore intended our results to be read as a *layer-resolved* probe of morphology in BERT, not as an equivalence between minimal-context BERT and a context average.

2 Results

2.1 fastText vs. BERT as target spaces

Table 1 reports `fastText` alongside BERT results taken from the *top (12th) layer*. The qualitative pattern replicates across spaces: starting from stem-only, adding *case* to the main-effects model yields the first substantial gain (33.01% for $\text{BERT}_{\ell=12}$; 35.7% for `fastText`); adding the *number×case* interaction improves further; and the four-way bundle (number:case:possessive:clitic) reaches the ceiling. The top-layer BERT ceiling is modestly lower than fastText (86.16% vs. 89%).

Table 2 contrasts error types (share of all errors). Relative to `fastText`, BERT (top, 12th layer) shows more *case* errors (35.3% vs. 3.7%), more *lexeme* exchanges (22.4% vs. 16.5%), and more *number* errors (17.5% vs. 9.9%), while reducing *clitic-only* errors (0.48% vs. 6.4%). Overabundance occupies a larger fraction for BERT (11.4% vs. 7.5%); excluding these raises BERT from 86.16% to ≈87.7% and `fastText` to ≈92%.

Model	fastText	BERT
Stem only	3.6	3.62
Stem + Number	7.0	7.45
Stem + Case	35.7	33.01
Stem + Number + Case + Poss + Clitic	75.6	75.13
+ Number:Case	82.4	81.87
+ Number:Case:Poss:Clitic	89.0	86.16

Table 1: Accuracies (%) of generating models: fastText (Nikolaev et al., 2022) vs. BERT (top, 12th layer; this study). Evaluation by best correlation with gold targets.

Error category	fastText	BERT
Case	3.7%	35.3%
Lexeme (stem exchange)	16.5%	22.4%
Number	9.9%	17.5%
Overabundance	7.5%	11.4%
Possessive	4.3%	4.6%
Clitic (alone)	6.4%	0.48%

Table 2: Top error categories (share of all errors): fastText (Nikolaev et al., 2022) vs. BERT (top, 12th layer; this study)

Both spaces reward the same interaction structure, supporting interaction-rich inflectional semantics. BERT's lower ceiling is driven by case/number confusions and more lexeme swaps, suggesting softer neighborhoods. Conversely, clitic-only errors are rarer with BERT, consistent with contextual localization of discourse particles.

2.2 Unsupervised structure of BERT noun embeddings

We visualized BERT embeddings (top, 12th layer) with t-SNE, coloring by case, number, possessive, and clitic (Figures 1–2). t-SNE preserves local neighborhoods rather than global axes.

Case yields visible macro-organization with semi-separated islands (e.g., locatives, PAR, GEN), but with broad overlap and diffuse borders, consistent with the need for interactions and residual case confusions.

Singular/plural show interdigitated strata with small pockets of separation; number is salient locally, but boundaries are porous.

Possessive marking forms localized patches (notably 2SG, 3SG), shaping local neighborhoods without dominating the global layout.

Clitic-bearing forms occupy small, compact pe-

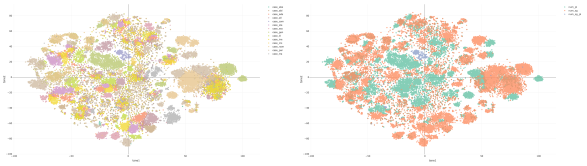

Figure 1: t-SNEs of BERT (top, 12th layer) noun embeddings: case (left) and number (right).

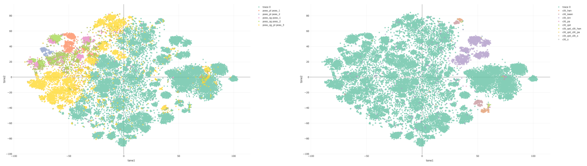

Figure 2: t-SNEs of BERT (top, 12th layer) noun embeddings: possessive (left) and clitic (right).

ripheral clusters; presence is encoded sharply when it occurs but is globally sparse, matching the low rate of clitic-only errors.

2.3 Layer-wise results for BERT noun embeddings

We evaluated the full generating model (main effects for stem, number, case, possessive, clitic, plus the number:case:possessive:clitic interaction) separately for each FinBERT layer $\ell \in \{1, \ldots, 12\}$ using the same inventory of 55,271 forms and the same evaluation protocol (rank-1 nearest correlation) as in Nikolaev et al., 2022.

Figure 3 summarizes the layer-wise accuracies. Accuracy rises steeply from the lowest layers to layer 4 and then varies within a narrow band until the top layer: L1 67.21%, L2 76.02%, L3 82.77%, L4 84.23%, L5 83.90%, L6 83.45%, L7 83.44%, L8 81.27%, L9 81.57%, L10 82.46%, L11 82.56%, and L12 86.16%. The best performance is at the top (12th) layer.

Figure 4 reports, for each layer, the within-layer composition of errors (shares summing to 100%). Clitic-related errors are rare at all depths, and overabundance contributes a stable minority of the error mass. The relative weighting of CASE and NUMBER varies with depth: compared to the top layer, several middle layers show a lower share of CASE errors and a higher share of NUMBER errors. Figure 5 makes this explicit by plotting, for each layer, the log-odds difference in the share of CASE and NUMBER errors relative to layer 12.

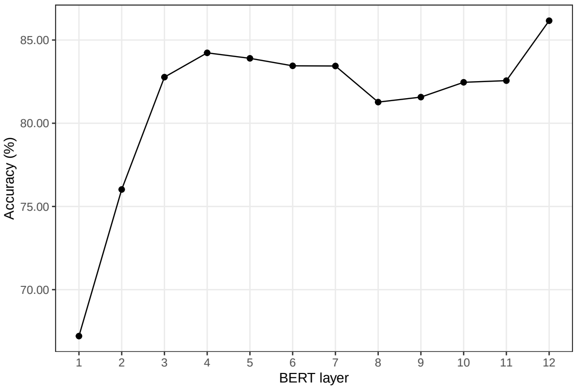

Figure 3: Accuracy by BERT layer (full model).

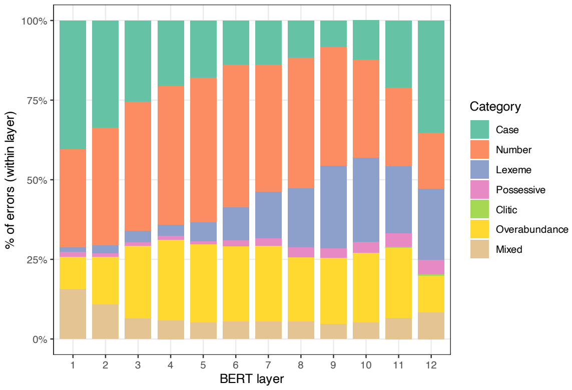

Figure 4: Within-layer error composition (shares sum to 100%) across layers 1–12.

3 Discussion

Using a fixed generating model, we find a *Case–Number dissociation* across BERT's depth: mid layers best support CASE (a lower share of case errors), the top layer best supports NUMBER (highest overall accuracy with a lower share of number errors), while fastText yields crisper case geometry and a slightly higher ceiling. We cast the comparison in layered terms (treating each FinBERT layer as its own target space) to ask where in the stack inflectional cues become recoverable. Two results are stable across all settings. First, inflectional meaning is *distributed and interaction-rich*: adding *case* to stem features yields the first major improvement, the *number*×*case* interaction adds a further jump, and a higher-order bundle (number:case:possessive:clitic) reaches the ceiling. Second, representation design and depth determine *which* cues are easiest to recover.

In our setup, fastText remains a morphology-forward baseline: character n-grams overlap suffixal material and produce crisp case geometry

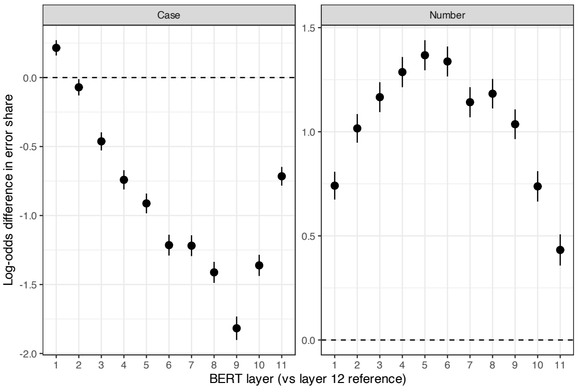

Figure 5: Log-odds difference in error share for Case and Number relative to layer 12 (95% CIs).

with a slightly higher ceiling. By contrast, our BERT targets are usage-trained but extracted with minimal context, and the layerwise pattern aligns with Booij's distinction between *inherent* vs. *contextual* inflection (Booij, 2012): Number is inherently chosen (a lexical–semantic property of the noun phrase), whereas Case is typically contextually assigned by government or agreement (a dependency with verbs, adpositions, or nominal heads). Without sentence context at extraction, case cues must be recovered from priors learned in pretraining. This helps explain the graded dissociation we observe: several middle layers (where morpho-syntactic regularities are strongest) show a *lower* share of Case errors but a *higher* share of Number errors, while the top (12th) layer (where broader lexico-semantic structure dominates) yields the best overall accuracy yet contributes a relatively larger share of residual Case errors and fewer Number errors. Clitic-only errors are rare at all depths, and possessive contributes a small, stable portion of the error mass.

A further depth effect concerns Lexeme-swap errors (predicting the right slot of the wrong lemma): these are small low in the stack but *increase toward the top*, consistent with a shift from form-anchored identity to lexico-semantic attraction as depth grows. This pattern fits with evidence that segmentation choices condition what morphology is recoverable in Transformer spaces: morphology-aware segmentations can improve performance and invite a dual-route view in which models sometimes store whole forms and sometimes compose them from parts (Hofmann et al., 2021). In our setting we kept WordPiece fixed and used minimal context, so the layerwise curves should be read as localizing *priors* learned in pretraining (not sentence-conditioned assignment at test time). Two concrete predictions follow for future work: averaging token vectors over diverse sentence contexts should attenuate lexeme competition, and adopting morpheme-aligned segmentation for Finnish should sharpen case recoverability. A Finnish-specific caveat is that pervasive consonant gradation and stem allomorphy mean that strictly morpheme-boundary tokenization can hide useful *boundary-spanning* cues: the very substrings that fastText's character n-grams exploit and that BERT may capture through sequences of WordPieces. We therefore expect hybrid interventions (morpheme-aligned units *plus* boundary-spanning character features, or explicit modeling of gradation/allomorphy) to outperform a purely morpheme-segmented vocabulary. The present layer-resolved map provides the baseline against which these Finnish-specific design choices can be measured.

References

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Geert Booij. 2012. *The grammar of words: An introduction to linguistic morphology*. Oxford University Press.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Valentin Hofmann, Janet B Pierrehumbert, and Hinrich Schütze. 2021. Superbizarre is not superb: Derivational morphology improves bert's interpretation of complex words. *arXiv preprint arXiv:2101.00403*.

Alexandre Nikolaev, Yu-Ying Chuang, and R Harald Baayen. 2022. A generating model for finnish nominal inflection using distributional semantics. *The Mental Lexicon*, 17(3):368–394.

Mike Schuster and Kaisuke Nakajima. 2012. Japanese and korean voice search. In *2012 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5149–5152. IEEE.

Antti Virtanen, Jenna Kanerva, Rami Ilo, Jouni Luoma, Juhani Luotolahti, Tapio Salakoski, Filip Ginter, and Sampo Pyysalo. 2019. Multilingual is not enough: Bert for finnish. *arXiv preprint arXiv:1912.07076*.