Multi-SimLex for Dutch: Benchmarking Embedding- and Prompt-Based Model Performance on Semantic Similarity

Lizzy Brans[♦], **Jelke Bloem**[♣♠]

♦ Human Computer Interaction, Utrecht University
♣ Institute for Logic, Language and Computation, University of Amsterdam
♠ Data Science Centre, University of Amsterdam
lizzybrans@hotmail.com, j.bloem@uva.nl

Abstract

We introduce Dutch Multi-SimLex, a 1,888–pair extension of the Multi-SimLex benchmark for evaluating lexical semantic similarity in Dutch. The dataset was rated by 100 native speakers on a 0–6 scale and shows high reliability (overall ICC(2,k)=0.82) as well as strong alignment with English (ρ=0.73). Using this resource, we evaluate eighteen models across four architectural families: static embeddings, encoder-only transformers, encoder–decoders, and decoder-only LLMs. We evaluate models using two complementary approaches: embedding-based cosine similarity and prompted similarity judgments in Dutch. In embedding-based evaluation, FastText (ρ=0.485) and the monolingual Dutch encoder BERTje (ρ=0.468) achieve the strongest alignment with human ratings, while multilingual encoders such as mBERT (ρ=0.208) and XLM-R (ρ=0.186) perform weaker. Prompt-based evaluation yields substantially higher correlations, with GPT-4 (ρ=0.761) performing best, followed by DeepSeek-V3 (ρ=0.753) and Gemini 1.5 Pro (ρ=0.722). Together, the results show that model performance depends strongly on how meaning is tested. Dutch Multi-SimLex provides a reliable foundation for evaluating meaning across architectures and advancing Dutch semantic evaluation.

Keywords: semantic similarity, Dutch, embeddings, prompting, evaluation

1. Introduction

Recent advances in large language models (LLMs), such as instruction-tuned generative decoder models, have made natural language interaction with computers increasingly human-like (Achiam et al., 2023; Radford et al., 2019). While their fluency is impressive, a key question remains: how do these models operationalise lexical meaning, and to what extent do their representations align with human semantic judgements? Without grounding in real-world experience, models risk confusing paradigmatic semantic similarity with syntagmatic association, e.g. rating *coffee* and *cup* as similar rather than related (Hill et al., 2015; Vulić et al., 2020). This distinction is crucial for reliable NLP applications such as retrieval, dialogue management, and text simplification.

Semantic similarity tasks provide a way to investigate this issue by measuring the degree to which models align with human judgments of lexical meaning (Miller and Charles, 1991). While early resources such as WordSim-353 or MEN conflated relatedness with similarity (Finkelstein et al., 2001; Bruni et al., 2014), SimLex-999 explicitly distinguished the two (Hill et al., 2015). Multi-SimLex extended this approach to twelve typologically diverse languages, using re-annotation by native speakers to ensure cross-lingual validity (Vulić et al., 2020).

Dutch poses a relevant test case as it is widely used in a highly digitised society, yet remains underrepresented in multilingual benchmarks and LLM training data (Odijk, 2012; Vanroy, 2023). While a Dutch SimLex-999 exists (Brans and Bloem, 2024), it lacks the scale and multilingual alignment of Multi-SimLex. This underrepresentation and the "curse of multilingualism" create challenges for evaluating Dutch semantic representations, as multilingual models often underperform in individual languages (Vlantis and Bloem, 2025).

This paper introduces a Dutch extension of Multi-SimLex and evaluates a broad spectrum of model types, from static embeddings (Word2Vec, FastText, GloVe) to contextual encoders (BERTje, RobBERT, mBERT, XLM-R, EuroBERT), encoder–decoder models (mT5, FLAN-T5), and decoder-only LLMs (BLOOM, Falcon-7B, Schaapje-2B, GPT-4, Gemini, DeepSeek-V3). We report results from both embedding-based and prompt-based approaches to obtain complementary insights, without treating the two paradigms as directly comparable.

We have four key contributions:

1. A Dutch extension of Multi-SimLex with 1,888 human-rated word pairs.

2. A cross-architectural evaluation of Dutch semantic similarity covering static, encoder, encoder–decoder, and decoder-only models.

3. An analysis of embedding-based and prompt-based evaluation methods, showing how each

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 4846–4860
11–16 May 2026.

highlights different aspects of semantic knowledge.

4. Empirical insights into how architectural design, multilingual scope, and language-specific adaptation affect representation of Dutch lexical semantics.

Overall, this study provides both a resource for Dutch NLP[1] and an empirical investigation of how different architectures encode meaning. It aims to guide future work towards more robust and linguistically informed evaluations in multilingual NLP.

2. Related Work

This section summarises prior research on model architectures and intrinsic evaluation strategies relevant to multilingual and Dutch NLP.

2.1. Model Architectures

Static word embeddings (Word2Vec, GloVe, FastText) operationalise the distributional hypothesis and have long been the default for intrinsic similarity evaluation (Mikolov et al., 2013; Pennington et al., 2014; Bojanowski et al., 2017). These models learn a single vector per word, independent of context, and capture general co-occurrence patterns rather than contextual nuances.

Contextual encoders such as BERT and RoBERTa, including Dutch adaptations BERTje and RobBERT/RobBERT-2023, improve on these tasks through masked language modelling and bidirectional self-attention (Devlin et al., 2019; Liu et al., 2019; de Vries et al., 2019; Delobelle et al., 2020; Delobelle and Remy, 2024). Encoder models process input sequences in parallel, building hierarchical representations where meaning emerges from contextual relations between tokens across all layers.

Multilingual encoders (mBERT, XLM-R, EuroBERT) extend this principle across languages, encoding shared subword vocabularies to promote cross-lingual transfer. However, this often comes at the cost of language-specific fidelity due to vocabulary and parameter sharing (Devlin et al., 2019; Conneau et al., 2020; Boizard et al., 2025).

Encoder–decoder models (T5, mT5, FLAN-T5) employ a sequence-to-sequence transformer architecture (Raffel et al., 2020; Xue et al., 2021; Chung et al., 2024), where the encoder captures input meaning and the decoder generates task-specific outputs. By framing all tasks in a text-to-text format, they can unify training objectives across classification, generation, and translation tasks, which makes them particularly suited for prompt-based evaluation.

Finally, decoder-only LLMs (e.g., GPT-4, LLaMA, BLOOM, Gemini, DeepSeek) consist solely of autoregressive transformer blocks that predict the next token in a sequence (Achiam et al., 2023; Touvron et al., 2023; Le Scao et al., 2023; Gemini Team, 2024; Liu et al., 2024). These models excel at text generation, simulated reasoning and instruction-following, particularly when instruction tuned, yet their internal embeddings are not always directly accessible or standardised across implementations. Their semantic representations must therefore be probed either via derived embeddings or through prompt-based tasks, as explored in this study.

2.2. Intrinsic Evaluation Approaches

Intrinsic evaluation measures how closely model-derived metrics align with human judgments. Models are intrinsically evaluated for grammatical capabilities through benchmarks of minimal pairs such as MultiBLIMP (Jumelet et al., 2026), in which a model should assign a higher probability to a grammatical variant of a sentence than an ungrammatical one. For the Dutch grammaticality benchmarks of the ANS (Pestel et al., 2025) and BLIMP-NL (Suijkerbuijk et al., 2025), results can also be compared to human acceptability judgements for the same sentences.

In the domain of semantic representations, intrinsic evaluations offer a controlled way to analyse how models represent meaning internally (Hill et al., 2015; Vulić et al., 2020). Unlike extrinsic evaluation, which tests downstream performance, intrinsic tasks isolate linguistic abilities and provide detailed insight into a model's semantic representations. Intrinsic evaluation of semantic representations is based on benchmarks of human word or sentence similarity ratings to which the similarity of model representations of those same words or sentences is compared.

Embedding-based methods assess similarity by computing cosine distances between word vectors (Rogers et al., 2020). Static embeddings such as Word2Vec, GloVe, and FastText assign one vector to each word regardless of context. Contextual models like BERT or XLM-R instead generate token-specific vectors that adapt to surrounding words. This approach captures the geometry of a model's semantic space. However, it assumes that spatial proximity fully represents meaning, which is problematic for words with multiple senses or strong contextual variation.

Prompt-based methods are increasingly used for large language models (Brown et al., 2020; Wei et al., 2022). Instead of analysing embeddings, prompting instruction-tuned models makes it possible to ask models to judge

[1] https://github.com/lizzybrans/Dutch-MultiSimLex

similarity directly through natural-language instructions (Snelder et al., in press). In a zero-shot setup, a model receives a prompt such as `"On a scale of 0–10, how similar are [word1] and [word2]?"`. Few-shot prompts include a few examples to guide the model, while Chain-of-Thought prompting encourages step-by-step reasoning before producing a score. This approach tests semantic reasoning rather than spatial encoding, but it is sensitive to prompt wording and less reproducible across models (Webson and Pavlick, 2022; Maia Polo et al., 2024). Prompting has been applied directly to semantic similarity tasks, showing that large language models can approximate human similarity judgments through zero-shot or few-shot instructions (Su et al., 2024; Kirk et al., 2023). However, results remain highly sensitive to linguistic phrasing and rating scale design, which highlights the need for standardised prompts in intrinsic evaluation (He et al., 2024; Zhu et al., 2024).

Researchers are also starting to explore the capabilities of instruction-tuned generative decoder models for replicating human linguistic judgements of specific lexical-semantic and psycholinguistically relevant properties. Brysbaert et al. (2024) perform zero-shot prompting for word familiarity estimates using GPT-4, observing similar correlations as between human annotators. Trott (2024) found that GPT-4 is quite good at generating word similarity scores that correlate with human judgements, reporting a correlation of 0.86 with the English SimLex-999 benchmark. De Deyne (2024) perform a different task to elicit similar judgements. Snelder et al. (in press) explore different prompting strategies for similarity judgements across different models and languages, noting that effective prompting strategies are quite model-dependent. They do observe that prompting for the judgements pair by pair is more effective, avoiding interference from other pairs in the context.

The two approaches are complementary. Embedding-based evaluation probes what a model encodes, while prompt-based evaluation examines how it applies that knowledge. Our evaluation combines both approaches to provide a more complete picture of how architecture and training shape Dutch semantic similarity.

3. Method

This section describes the construction of the Dutch Multi-SimLex dataset, the models evaluated, and the intrinsic evaluation protocols.

3.1. Dutch Multi–SimLex Construction

We extend Multi–SimLex to Dutch following the original annotation and translation protocol (Vulić et al., 2020). Half of the 1,888 word pairs were taken from the existing Dutch SimLex-999 and rescaled from a 0–10 to a 0–6 range (Brans and Bloem, 2024).[2] This and other deviations from the standard Multi-SimLex protocol were approved by the authors of Multi-SimLex (Vulić, 2024). The remaining pairs were newly translated into Dutch following the Multi–SimLex guidelines: (i) pairs must be unique; (ii) English synonyms may not collapse into the same Dutch form; (iii) translations must preserve the original semantic relation; (iv) multi-word expressions are permitted when no single-word equivalent exists (Vulić et al., 2020).

Annotation protocol. To manage the annotator workload, the dataset was divided into nine subsets of 100 pairs, a deviation from the original protocol. Each subset was rated by at least ten native Dutch speakers on a continuous 0–6 similarity scale. Word pairs were presented in random order. Participants were aged 16 or older, gave informed consent, and completed the survey online. The instructions followed the SimLex-999 guidelines (Hill et al., 2015), emphasising similarity rather than association (e.g., *kop–mok* vs. *auto–snelweg*). Our Dutch version of the instructions is shown in Appendix D for reproducibility.

Post-processing and reliability. We removed outlier raters whose scores deviated strongly from group patterns. Pairs with high variance were checked for ambiguity. Final similarity scores were computed as the mean per pair. Inter-annotator agreement was measured with the two-way random-effects Intraclass Correlation Coefficient (ICC), reported for individual raters (ICC(2,1)) and averaged ratings (ICC(2,k)).

In this framework, ICC(2,1) estimates the reliability of a single rater, whereas ICC(2,k) estimates the reliability of the mean rating across k raters. Formally, ICC(2,1) is defined as:

$$\mathrm{ICC}(2,1) = \frac{MS_B - MS_E}{MS_B + (k-1)MS_E + \frac{k}{n}(MS_R - MS_E)}$$

where MS_B is the between-target mean square, MS_R the between-rater mean square, MS_E the residual mean square, k the number of raters, and n the number of targets. ICC(2,k) is derived as:

$$\mathrm{ICC}(2,k) = \frac{MS_B - MS_E}{MS_B + \frac{1}{n}(MS_R - MS_E)},$$

[2] To verify that the rescaling preserved the original semantic judgements, we compared the rescaled Dutch SimLex-999 scores with the original ratings. Because the transformation is linear, the rank ordering of the pairs remains unchanged, and therefore Spearman's ρ is preserved.

reflecting the increased stability obtained by averaging across raters.

Dutch Multi-SimLex is available on GitHub: https://github.com/lizzybrans/Dutch-MultiSimLex.

3.2. Models Evaluated

We evaluated four model families to capture both static and contextual architectures:

- **Static embeddings:** Word2Vec, GloVe, and FastText trained on Dutch OSCAR (Mikolov et al., 2013; Pennington et al., 2014; Bojanowski et al., 2017).
- **Encoders:** Dutch BERTje and RobBERT (v1, v2023), and multilingual models mBERT, XLM-R, and EuroBERT (210M, 2.1B) (de Vries et al., 2019; Delobelle et al., 2020; Delobelle and Remy, 2024; Devlin et al., 2019; Conneau et al., 2020; Boizard et al., 2025).
- **Encoder–decoders:** mT5 and FLAN-T5, which process text in a sequence-to-sequence format (Xue et al., 2021; Chung et al., 2024).
- **Decoders:** BLOOM, Falcon-7B, and Schaapje-2B (open-weight) via embeddings; GPT-4, Gemini 1.5 Pro, and DeepSeek-V3 via prompting due to API-only access (Le Scao et al., 2023; Almazrouei et al., 2023; Smits, 2024; Achiam et al., 2023; Gemini Team, 2024; Liu et al., 2024). Although BLOOM is trained on 46 languages, its dataset does not include Dutch, making its evaluation effectively zero-shot for Dutch.

These model families represent the main Transformer configurations. Encoder models use bidirectional attention to learn contextual word representations. Encoder–decoder models handle text generation and comprehension through separate encoding and decoding stages. Decoder-only models predict the next token in a sequence, making them effective generators but harder to probe through embeddings. Detailed technical specifications for all evaluated models, including parameter sizes, training corpora, and tokenizers, are provided in Appendix B.

3.3. Evaluation Protocols

3.3.1. Embedding-based evaluation

For each word w, we obtain a vector $\mathbf{e}(w)$. Static models provide pre-computed lookup vectors.

For transformers, we extract hidden states from the final layer and compute their mean across subword tokens. This isolated encoding setup follows standard intrinsic similarity evaluations in which contextual models are probed without additional surrounding text to assess the stability of their lexical representations (Rogers et al., 2020). Contextualised representations can alternatively be reduced to static embeddings by averaging across diverse sentential contexts (Bommasani et al., 2020). The present design instead compares architectures under a controlled and uniform probing condition, complemented by a joint-encoding variant that allows limited contextual interaction within each word pair.

For a pair (w_1, w_2), cosine similarity is defined as:

$$\mathrm{sim}(w_1, w_2) = \frac{\mathbf{e}(w_1) \cdot \mathbf{e}(w_2)}{\|\mathbf{e}(w_1)\| \, \|\mathbf{e}(w_2)\|}.$$

Spearman's rank correlation (ρ) between model and human ratings measures how well the model captures similarity. To mirror prompting conditions, we also apply a joint-encoding variant. In this setup, "w_1 w_2" is encoded together, and cosine similarity is computed between their pooled representations. For sequence-to-sequence models, we only take embeddings from the encoder part of the architecture. Unless noted otherwise, we report final-layer results following Rogers et al. (2020). Out-of-vocabulary (OOV) cases in static models without subtokenization are excluded, and effective sample sizes are reported.

3.3.2. Prompt-based evaluation

Generative models are queried in Dutch using three templates aligned to the 0–6 scale:
(i) zero-shot (instruction only)

```
Op een schaal van 0 tot 6, hoe
vergelijkbaar zijn de woorden
[woord1] en [woord2]?
Antwoord alleen met het numerieke
cijfer.
```

(ii) few-shot (three labelled Dutch examples)

```
Kop en mok lijken op elkaar (6).
Alligator en krokodil lijken
enigszins op elkaar (3).
Auto en snelweg lijken niet op
elkaar (0).
Gegeven dit patroon:  op een
schaal van 0 tot 6, hoe vergeli-
jkbaar zijn [woord1] en [woord2]?
Antwoord alleen met het numerieke
cijfer.
```

(iii) chain-of-thought (a brief reasoning step before scoring) (Brown et al., 2020; Wei et al., 2022).

```
Maak eerst een lijst van se-
mantische kenmerken die worden
gedeeld door [woord1] en [wo-
ord2].
Beoordeel daarna, op een schaal
van 0 tot 6, hoe vergelijkbaar de
```

```
woorden zijn, waarbij:
- 0.0 = geen gedeelde kenmerken
- 3.0 = enige semantische overlap
- 6.0 = identieke betekenis
Antwoord alleen met het numerieke
cijfer.
```

Test items are presented individually, not in batches, as this leads to better performance (Snelder et al., in press). Prompts require numeric-only answers. If a model produces additional text, the first number between 0 and 6 is extracted automatically using a regular expression. All values are then normalised to a consistent 0–6 range. We compute Spearman's ρ between the model outputs and human similarity ratings. Templates are kept identical across models to maintain comparability and reduce prompt variance (Webson and Pavlick, 2022; Maia Polo et al., 2024).

3.3.3. Implementation

All open-weight models are run locally with default tokenisers. Multi-word expressions are space-separated. FastText handles OOV words through subword composition; other static models omit OOV pairs, and we report the reduced sample size. Encoder hidden states are obtained using `output_hidden_states=True` and mean-pooled across subtoken spans. API-based evaluations for GPT-4, Gemini, and DeepSeek use identical Dutch prompts and the same parsing script. Dataset splits, annotation materials, and evaluation code will be released publicly for replication.

4. Results

This section presents the validation of the Dutch Multi-SimLex dataset and the evaluation results across model architectures. All results are reported as Spearman's ρ correlations between model predictions and human similarity judgments on the 0–6 scale.

4.1. Dataset Validation and Multilingual Alignment

After quality control, 100 valid Dutch annotators completed the survey. Quality control consisted of filtering out participants reporting language disorders, and participants whose ratings deviated substantially from the group mean. There was a pilot study with 10 participants in which the optimal number of pairs per page was determined (100 on one page rather than 50 or 25 pairs on multiple pages) and where words with diverging ratings were flagged for potential issues in the translation.

The 1,888 word pairs were divided into nine subsets of approximately 100 pairs each, with an average of 11 raters per subset. Inter-rater agreement was high, with ICC(2,k) values ranging from 0.84 to 0.95 across subsets and an overall ICC(2,k) of 0.82 [0.74–0.89] (Table 1). These results indicate excellent reliability for averaged ratings, comparable to or higher than the values reported in other Multi-SimLex languages (Vulić et al., 2020). The lower ICC(2,1) value (0.25) reflects the expected variability of individual raters on a fine-grained similarity task, whereas ICC(2,k) increases substantially because averaging across approximately 11 raters per subset reduces random error and stabilises the aggregate score.

Measure	ICC	95% CI
ICC(2,1)	0.25	[0.17, 0.36]
ICC(2,k)	0.82	[0.74, 0.89]

Table 1: Overall inter-rater reliability across Dutch annotators.

Cross-lingual correlations show that Dutch aligns most strongly with English (ρ = 0.73), Welsh, and French (both 0.70), and least with typologically distant languages such as Swahili (0.59) (Table 2). This pattern is consistent with previously reported cross-lingual correlations in Multi-SimLex, where European languages tend to show stronger mutual alignment than typologically distant languages (Vulić et al., 2020). The results confirm that the Dutch data integrate consistently into the multilingual benchmark.

Some pairs revealed cultural or lexical ambiguity (e.g., *kerk–tempel*, *fiets–motor*), but these were retained to maintain semantic diversity and represent the breadth of Dutch conceptual categories.

Language	ρ
English (ENG)	0.73
Welsh (CYM)	0.70
French (FRA)	0.70
Russian (RUS)	0.69
Estonian (EST)	0.69
Spanish (SPA)	0.69
Polish (POL)	0.68
Finnish (FIN)	0.67
Hebrew (HEB)	0.67
Mandarin (CMN)	0.66
Arabic (ARA)	0.65
Cantonese (YUE)	0.63
Swahili (SWA)	0.59

Table 2: Top- and bottom-aligned languages with Dutch. Full results are provided in the Appendix.

Across all nine annotation groups, some items elicited substantial disagreement, indicating semantic ambiguity or context effects. Table 3 lists the five pairs with the highest standard deviation (SD) in human scores. Notably, all are nouns suggest-

ing that conceptual richness and polysemy drive variance even in a similarity-focused task. These descriptives complement the strong ICC values and help contextualise model–human mismatches.

Word 1	Word 2	SD
winnaar *winner*	aanwezigheid *presence*	2.51
bowlen *bowling*	werper *bowler*	2.35
verjaardag *birthday*	datum *date*	2.34
klooster *monastery*	douche *shower*	2.31
zuiger *piston*	aandrijving *engine*	2.22

Table 3: Pairs with highest variance in Dutch human ratings (SD). Provided translations are the corresponding English word pairs.

4.2. Embedding-based Model Performance

Table 4 summarises embedding-based results for both the isolated condition (the word is embedded without context) the and joint encoding condition (the word is embedded with the other word of the pair as context). Results are reported separately for each variant in order to assess the extent to which limited contextual interaction between paired words affects correlation with human similarity judgements.

4.2.1. Static embeddings

Among static models, we see in Table 4 that FastText achieves the highest correlation (ρ = 0.485; OOV-aware = 0.504), followed by Word2Vec (0.431) and GloVe (0.360 at 100K sentences). The advantage of FastText likely stems from its subword-based architecture, which models Dutch morphology more effectively than word-level embeddings (Remy et al., 2023). For example, morphologically related forms such as *werk–werker* ('work–worker') or Dutch compounds like *fiets–fietsbel* ('bicycle–bicycle bell') share character n-grams that remain visible to FastText's subword representation. This design can yield more stable similarity estimates for morphologically complex words compared to token-based embeddings. This result aligns with previous findings that subword composition mitigates sparsity in morphologically rich languages (Bojanowski et al., 2017). Overall, these results show that even relatively simple architectures can produce strong representations when trained on large, language-specific corpora.

Model	Isolated ρ	Joint ρ
FastText	**0.485**	N/A
Word2Vec	0.431	N/A
GloVe (100k)	0.360	N/A
BERTje (NL)	0.440 (L1)	**0.468** (L2)
RobBERT (NL)	0.203 (L5)	0.355 (L4)
RobBERT-2023(NL)	0.398 (L0)	0.439 (L0)
mBERT	0.158 (L1)	0.208 (L7)
XLM-R	0.179 (L0)	0.186 (L0)
EuroBERT-210M	0.078 (L3)	0.054 (L3)
EuroBERT-2.1B	0.051 (L0)	0.056 (L1)
mT5	0.096 (L0)	**0.153** (L0)
FLAN-T5	0.062 (L0)	0.069 (L12)
Schaapje-2B (NL)	0.042 (L2)	**0.091** (L12)
BLOOM-560m	0.037 (L17)	0.080 (L4)
Falcon-7B	0.080 (L3)	0.055 (L2)

Table 4: Embedding-based Spearman ρ for isolated and joint encodings on Dutch Multi–SimLex.

4.2.2. Encoder-only models

Among contextual models, Table 4 shows that Dutch-specific encoders outperform multilingual ones. BERTje achieves the strongest correlations (embedding L1 = 0.440; contextualised L2 = 0.468), while RobBERT improves when words are embedded with their paired word as context (contextualized condition, 0.203 → 0.355). RobBERT-2023 reaches its best performance at the embedding level (0.398) with similar contextual trends (0.439). These findings highlight the benefits of domain-specific pre-training and vocabulary adaptation.

Multilingual encoders, including mBERT and XLM-R, perform notably lower (best ρ = 0.208 and 0.186, respectively), reflecting the trade-off between multilingual scope and language-specific depth. EuroBERT (210M / 2.1B) performs weakest (≤ 0.08), suggesting that extensive multilingual capacity does not compensate for limited Dutch token coverage.

Taken together, the encoder results demonstrate that pre-training on high-quality Dutch data remains more beneficial than scaling model size or multilingual reach. Across encoder models, peak correlations usually appear in the first few layers, where general lexical meaning is represented before deeper contextualisation occurs. We report both L0 and the best-performing layer to capture this trend without overemphasising individual layer peaks. For further details on layer-wise behaviour, Figure 1 in Appendix C reports correlations with human ratings across all layers. Table 4 reports the best-performing layer for each model, though this should be viewed as an empirical upper bound rather than a fixed evaluation protocol.

4.2.3. Encoder–decoders

Models such as mT5 and FLAN-T5 show consistently low correlations in Table 4 (best ρ = 0.153 and 0.069), even though their architectures are more flexible. Full layer-wise results are shown in Figure 2 in Appendix C.

This supports earlier observations that sequence-to-sequence transformers, optimised for generative objectives, tend to underperform on static similarity tasks (Raffel et al., 2020; Xue et al., 2021). Their dual-stage structure may diffuse lexical representations across encoder and decoder layers, reducing precision in word-level similarity evaluation.

4.2.4. Decoder-only models

Open-weight decoder models, including BLOOM, Falcon-7B, and Schaapje-2B, also perform weakly (≤ 0.09, Table 4). For BLOOM, this result is expected since the model was not trained on Dutch data and is effectively evaluated in a zero-shot setting. Full layer-wise results are available in Figure 3 in Appendix C.

These architectures are primarily optimised for text generation and long-range coherence rather than stable lexical mapping. Their internal embeddings are less consistent across contexts, which likely explains their lower intrinsic similarity scores. This underscores the limitation of using autoregressive decoders for word-level semantic probing, despite their generative strengths.

4.3. Prompt-based Model Performance

Prompt-based evaluation offers a complementary perspective. Instruction-tuned LLMs queried directly in Dutch yield substantially higher correlations (Table 5). GPT-4 achieves ρ = 0.761 in the zero-shot setting, followed by DeepSeek-V3 (0.753) and Gemini 1.5 Pro (0.722). Few-shot and Chain-of-Thought prompting produce similar or slightly lower values.

These results suggest that instruction tuning and large-scale alignment improve a model's ability to verbalise semantic relations, even when internal embeddings are not explicitly structured for similarity. However, qualitative inspection shows that these models sometimes conflate syntagmatic associations (e.g., *oorlog–leger*) with paradigmatic semantic similarity, indicating that semantic reasoning remains partly associative.

4.3.1. Error patterns

To assess rating stability, we compute the Mean Absolute Error (MAE) as the average deviation between model predictions and Dutch Multi-SimLex scores. Lower values indicate more consistent similarity judgments relative to the human gold standard. As shown in Table 6, Gemini's zero-shot prompt produced the most stable results (MAE = 1.25). However, the most consistent prompt type differed across models: zero-shot yielded the lowest errors for Gemini and DeepSeek V3, whereas GPT-4 achieved its best stability under the CoT prompt. These differences suggest that prompt robustness is model-dependent, supporting previous findings on the variability of prompt-based reasoning (Webson and Pavlick, 2022; Snelder et al., in press).

Model	0-shot	Few-shot	CoT
GPT-4	0.761	0.749	0.635
Gemini 1.5 Pro	0.722	0.703	0.720
DeepSeek V3	0.753	0.693	0.707

Table 5: Prompt-based Spearman ρ for instruction-tuned decoder models.

Model	0-shot	Few-shot	CoT
GPT-4	1.44	1.58	1.38
Gemini 1.5 Pro	1.25	2.05	1.70
DeepSeek V3	1.45	1.68	1.56

Table 6: Prompt-based Mean Absolute Error (MAE) for decoder-only models on Dutch Multi–SimLex. Lower values indicate more stable similarity ratings relative to human judgments.

A closer inspection of model outputs reveals a consistent tendency to conflate syntagmatic association with paradigmatic semantic similarity, a well-known limitation in distributional representations (Hill et al., 2015; Vulić et al., 2020). For example, pairs such as *oorlog–leger* ('war–army') and *arts–ziekenhuis* ('doctor–hospital') received high similarity scores from several models despite low human ratings. Conversely, agreement was strongest for straightforward synonymic or hierarchical relations (e.g., *eten–voedsel*, 'food–eat' and *rivier–water*, 'river–water'). These findings indicate that while prompt-based models achieve higher overall correlations, they remain sensitive to prompt format and continue to overrate associative or thematically related pairs. Table 7 illustrates the largest and smallest divergences.

4.4. Comparison

Across all settings, FastText and Dutch-specific encoders achieve the strongest embedding-based correlations, while multilingual and encoder–decoder models lag behind. Decoder-only models provide weak static representations, but instruction-tuned LLMs yield the highest alignment when prompted directly in Dutch.

Model	Pair	$\lvert\Delta\rvert$
GPT-4 (Few)	klooster-douche	4.9
Gemini (Zero)	oorlog-leger	4.7
DeepSeek (Zero)	arts-ziekenhuis	4.6
GPT-4 (CoT)	boek-papier	0.4
Gemini (CoT)	eten-voedsel	0.1
DeepSeek (CoT)	rivier-water	0.1

Table 7: Examples of prompt-based model outputs compared to Dutch Multi–SimLex scores. Top: largest divergences (relatedness misinterpreted as similarity); bottom: closest alignments with human ratings.

These results highlight the complementarity between embedding-based and prompt-based evaluation: embedding-based methods probe a model's internal structure, whereas prompting reveals how this knowledge is applied in reasoning. Together, they suggest that while contextual encoders capture stable lexical geometry, large instruction-tuned LLMs approximate human-like judgments through learned reasoning rather than explicit semantic grounding.

5. Discussion

5.1. Summary of findings

The Dutch extension of Multi-SimLex yields reliable human judgements (overall ICC(2,k) = 0.82) and strong alignment with English ($\rho = 0.73$), confirming its suitability as an intrinsic benchmark for Dutch lexical semantics (Vulić et al., 2020; Hill et al., 2015). Model results reveal a clear divide between evaluation paradigms. In embedding-based evaluation, Dutch-specific encoders (BERTje, RobBERT-2023) and static embeddings (FastText) perform best. In contrast, prompt-based evaluation shows much higher alignment with human ratings, with instruction-tuned decoder models (GPT-4, Gemini, DeepSeek-V3) substantially outperforming all embedding-based approaches. Beyond rank alignment, we also observe model-dependent stability differences in prompt settings (MAE). Gemini is most stable in zero-shot, whereas GPT-4 is most stable with CoT, indicating that the "best" prompt varies by model and should be calibrated rather than assumed.

5.1.1. Embedding-based evaluation

FastText achieves the highest static correlation, benefiting from subword modelling that captures Dutch morphology and compounding (Bojanowski et al., 2017). Among encoders, BERTje and RobBERT-2023 outperform multilingual models such as mBERT, XLM-R, and EuroBERT, which show limited Dutch coverage and weaker lexical precision. Layer-wise inspection confirms that early layers encode the most stable lexical information, while deeper layers increasingly specialise for syntax (Rogers et al., 2020). These findings reaffirm that monolingual pre-training and linguistically aligned tokenisation remain essential for accurate Dutch semantic modelling (de Vries et al., 2019; Delobelle et al., 2020; Delobelle and Remy, 2024). By contrast, multilingual encoders display the expected "curse of multilingualism," where capacity sharing across languages dilutes per-language fidelity (Conneau et al., 2020).

5.1.2. Prompt-based evaluation

Prompted decoder models achieve the highest overall correlations. Zero-shot prompting in Dutch yields the strongest results, supporting the view that instruction-tuned LLMs store semantic knowledge in a procedural form accessible through natural language queries (Brown et al., 2020; Wei et al., 2022). Few-shot and Chain-of-Thought variants do not consistently improve performance, reflecting model-specific sensitivities rather than a single optimal prompt design. Despite their impressive alignment, qualitative analysis shows that LLMs still conflate topical relatedness with genuine similarity (e.g., *oorlog-leger*), echoing earlier challenges in intrinsic evaluation (Hill et al., 2015). This indicates that LLM reasoning may reproduce associative rather than paradigmatic semantic relations. Closer inspection shows that the largest differences between model and human ratings occur in thematic or associative pairs such as *arts–ziekenhuis* ('doctor–hospital'), while agreement is highest for synonymic or hierarchical relations like *eten–voedsel*. This pattern suggests that even advanced instruction-tuned models tend to reason associatively, despite producing high overall correlations.

5.2. Implications and future work

These findings highlight several directions for future research. First, the strong performance of BERTje, RobBERT-2023, and FastText underlines the importance of Dutch-specific modelling and tokenisation. Improving subword segmentation and continuing pre-training on Dutch corpora may further enhance lexical precision, especially for morphologically complex forms. Because most lexical information is encoded in the lower layers, targeted adaptation or lightweight fine-tuning at these layers may yield better cost–performance trade-offs than large-scale retraining.

Second, for generative models, prompt-based evaluation remains a promising but unstable diagnostic. Structured prompts, definition-based similarity tasks, or automatic calibration methods could

reduce prompt variance while maintaining interpretability (Brown et al., 2020; Wei et al., 2022). Combining prompt- and embedding-based metrics may also offer a more complete view of how models encode and apply meaning. Reporting both correlation and error-based metrics, such as MAE, and releasing per-model effective sample sizes and confidence intervals would make intrinsic evaluations more transparent and comparable across studies.

Finally, multilingual encoders could benefit from adapter-based fine-tuning or concept-alignment objectives using resources like Multi-SimLex (Conneau et al., 2020; Vulić et al., 2020). Extending the dataset with multi-word expressions, specialised vocabulary, and regional varieties would improve ecological validity and support broader evaluation of Dutch semantics in multilingual NLP. Cross-lingual comparisons further show that Dutch aligns most strongly with English, Welsh, and French, particularly for nouns. Expanding coverage for verbs, adjectives, and adverbs could therefore strengthen diagnostic value and support cross-linguistic generalisation.

6. Conclusion

We introduced a Dutch extension of Multi-SimLex and validated its reliability and cross-lingual alignment. Using this benchmark, we compared static embeddings, encoder-only, encoder–decoder, and decoder-only architectures across two intrinsic paradigms. The results reveal a clear divide between representational and procedural evaluations of meaning.

In embedding-based evaluation, Dutch-specific encoders (BERTje, RobBERT-2023) and FastText yield the strongest lexical similarity signals, particularly in early layers where static and morphological information is most stable. Multilingual encoders lag behind on Dutch due to capacity sharing and vocabulary dilution, confirming known trade-offs in multilingual representation learning (Conneau et al., 2020). In contrast, prompt-based evaluation highlights the procedural nature of large decoder models. Instruction-tuned LLMs such as GPT-4, Gemini, and DeepSeek achieve much higher alignment with human ratings, but remain sensitive to prompt phrasing and prone to confusing relatedness with paradigmatic semantic similarity.

Together, these findings show that "semantic competence" is not a fixed property but depends on how a model is probed: embeddings capture representational geometry, while prompts reveal context-sensitive reasoning. In practice, this means that monolingual encoders and subword-based models are well-suited for retrieval and lexical comparison tasks in Dutch, while instruction-tuned LLMs perform best for interactive or interpretive applications that rely on explicit reasoning.

The Dutch Multi-SimLex dataset and accompanying baselines provide a reproducible foundation for evaluating meaning representations in Dutch NLP. Future work should expand lexical coverage, refine prompt-based probing methods, and explore Dutch-targeted adaptation not only for multilingual encoders (Vlantis and Bloem, 2025), but also for decoders. These directions can strengthen the fairness and linguistic fidelity of semantic evaluation, advancing multilingual NLP toward models that are not only fluent but also semantically grounded.

7. Limitations

The results should be interpreted with several limitations in mind. The embedding-based setup favours models with stable lexical representations, such as static and encoder-only architectures. Generative models, including encoder–decoders and decoder-only LLMs, distribute meaning across parameters rather than in a single embedding space, which makes direct cosine comparison less representative (Raffel et al., 2020; Brown et al., 2020). Although the joint-encoding variant reduces this gap, its results are not directly comparable to prompt-based scores that rely on instruction-following and reasoning.

Layer selection and tokenisation also affect outcomes. We report the best-performing layer per model, which provides an empirical upper bound but introduces selection bias (Rogers et al., 2020). Differences in tokenisation schemes and Dutch morphology further influence results, as segmentation quality may partly explain model gaps.

Prompt-based evaluations add variability due to prompt phrasing, numeric parsing, and API constraints. Minor wording changes can alter model responses, and proprietary systems such as GPT-4, Gemini, and DeepSeek cannot be fully reproduced. Finally, Dutch Multi-SimLex, though reliable (overall ICC(2,k)=0.82), includes a limited set of 1,888 pairs and underrepresents multi-word expressions and culturally specific terms. This restricts both semantic coverage and the attainable ceiling for correlation with human ratings.

8. Ethics Statement

Human annotation. All participants were native speakers of Dutch aged 16 or older. Participation was voluntary and anonymous, and each participant gave informed consent before starting. The task involved rating everyday Dutch word pairs on a 0–6 similarity scale. Instructions followed the original SimLex-999 guidelines, emphasising similarity rather than association. No identifying or sensitive data were collected.

Compensation. Participants were not financially compensated. The study was designed as a short, low-risk online task and did not involve any form of deception or personal data collection.

Data governance. The Dutch Multi–SimLex dataset has been released for research use under an open licence, together with annotation materials and evaluation code. All data are stored in aggregate form without any personal identifiers. The resource follows the ethical and reproducibility standards set by the Multi–SimLex project.

Model use and reproducibility. Open-weight models were evaluated locally with default tokenisers. Proprietary systems (GPT-4, Gemini, DeepSeek-V3) were queried through identical Dutch prompts and processed with the same numeric parsing script. While results are replicable for open models, replication of closed models may vary across API versions.

Environmental considerations. All experiments were conducted with single-pass evaluations on existing models. No re-training or parameter tuning was performed to minimise computational cost.

9. Bibliographical References

Josh Achiam, Steven Adler, Sandhini Agarwal, et al. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, et al. 2023. The Falcon series of open language models. *arXiv preprint arXiv:2311.16867*.

Nicolas Boizard, Hippolyte Gisserot-Boukhlef, Duarte M Alves, André FT Martins, Ayoub Hammal, Caio Corro, Céline Hudelot, Emmanuel Malherbe, Etienne Malaboeuf, Fanny Jourdan, et al. 2025. EuroBERT: Scaling multilingual encoders for European languages. In *Second Conference on Language Modeling*.

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.

Rishi Bommasani, Kelly Davis, and Claire Cardie. 2020. Interpreting pretrained contextualized representations via reductions to static embeddings. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4758–4781, Online. Association for Computational Linguistics.

Lizzy Brans and Jelke Bloem. 2024. SimLex-999 for Dutch. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 14832–14845, Torino, Italia. ELRA and ICCL.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. In *Advances in neural information processing systems*, volume 33, pages 1877–1901.

Elia Bruni, Nam-Khanh Tran, and Marco Baroni. 2014. Multimodal distributional semantics. *Journal of Artificial Intelligence Research*, 49:1–47.

Marc Brysbaert, Gonzalo Martínez, and Pedro Reviriego. 2024. Moving beyond word frequency based on tally counting: AI-generated familiarity estimates of words and phrases are an interesting additional index of language knowledge. *Behavior Research Methods*, 57(1):28.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Simon De Deyne. 2024. Evaluating human-like similarity biases at every scale in large language models: Evidence from remote and basic-level triads. In *Proceedings of the annual meeting of the cognitive science society*, volume 46.

Wietse de Vries, Andreas van Cranenburgh, Arianna Bisazza, Malvina Nissim, Gertjan van Noord, and Gosse Bouma. 2019. BERTje: A Dutch BERT model. *arXiv preprint arXiv:1912.09582*.

Pieter Delobelle and François Remy. 2024. RobBERT-2023: Keeping Dutch language models up-to-date at a lower cost thanks to model conversion. *Computational Linguistics in the Netherlands Journal*, 13:193–203.

Pieter Delobelle, Thomas Winters, and Bettina Berendt. 2020. RobBERT: a Dutch RoBERTa-based Language Model. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3255–3265, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186. Association for Computational Linguistics.

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2001. Placing search in context: The concept revisited. In *Proceedings of the 10th International Conference on World Wide Web*, pages 406–414. ACM.

Gemini Team. 2024. Gemini 1.5: Technical report. Google DeepMind technical report.

Jia He, Mukund Rungta, David Koleczek, Arshdeep Sekhon, Franklin X Wang, and Sadid Hasan. 2024. Does prompt formatting have any impact on LLM performance? *arXiv preprint arXiv:2411.10541*.

Felix Hill, Roi Reichart, and Anna Korhonen. 2015. SimLex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4):665–695.

Jaap Jumelet, Leonie Weissweiler, Joakim Nivre, and Arianna Bisazza. 2026. MultiBLiMP 1.0: A massively multilingual benchmark of linguistic minimal pairs. *Transactions of the Association for Computational Linguistics*, 14:193–216.

Robert Kirk, Amy Zhang, Edward Grefenstette, and Tim Rocktäschel. 2023. A survey of zero-shot generalisation in deep reinforcement learning. *Journal of Artificial Intelligence Research*, 76:201–264.

Teven Le Scao, Angela Fan, Christopher Akiki, et al. 2023. BLOOM: A 176B-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Andy Liu, Bo Feng, Bowen Xue, et al. 2024. DeepSeek-V3 technical report. *arXiv preprint arXiv:2412.19437*.

Yinhan Liu, Myle Ott, Naman Goyal, et al. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Felipe Maia Polo, Ronald Xu, Lucas Weber, Mírian Silva, Onkar Bhardwaj, Leshem Choshen, Allysson de Oliveira, Yuekai Sun, and Mikhail Yurochkin. 2024. Efficient multi-prompt evaluation of LLMs. *Advances in Neural Information Processing Systems*, 37:22483–22512.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *Proceedings of ICLR (Workshop Track)*.

George A. Miller and Walter G. Charles. 1991. Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1):1–28.

Jan Odijk. 2012. *The Dutch Language in the Digital Age / De Nederlandse Taal in het Digitale Tijdperk*. META-NET White Paper Series. Springer, Berlin.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Julia Pestel, Jelke Bloem, and Raquel G Alhama. 2025. Evaluating Dutch speakers and large language models on standard Dutch: A grammatical challenge set based on the Algemene Nederlandse Spraakkunst. *Computational Linguistics in the Netherlands Journal*, 14:555–582.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

François Remy, Pieter Delobelle, Bettina Berendt, Kris Demuynck, and Thomas Demeester. 2023. Tik-to-tok: Translating language models one token at a time: An embedding initialization strategy for efficient language adaptation. *arXiv preprint arXiv:2310.03477*.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Robin Smits. 2024. Schaapje-2b-pretrained-gguf. Accessed: 2025-04-28.

Xander Snelder, Yunchong Huang, and Jelke Bloem. in press. Prompting instruction-tuned

LLMs for semantic similarity values. In *Proceedings of the Fifteenth International Conference on Language Resources and Evaluation (LREC 2026)*.

Yi Su, Yunpeng Tai, Yixin Ji, Juntao Li, Yan Bowen, and Min Zhang. 2024. Demonstration augmentation for zero-shot in-context learning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14232–14244, Bangkok, Thailand. Association for Computational Linguistics.

Michelle Suijkerbuijk, Zoë Prins, Marianne de Heer Kloots, Willem Zuidema, and Stefan L. Frank. 2025. BLiMP-NL: A corpus of Dutch minimal pairs and acceptability judgments for language model evaluation. *Computational Linguistics*, 51(4):1267–1301.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Sean Trott. 2024. Can large language models help augment english psycholinguistic datasets? *Behavior Research Methods*, 56(6):6082–6100.

Bram Vanroy. 2023. Language resources for dutch large language modelling.

Daniel Vlantis and Jelke Bloem. 2025. Intrinsic evaluation of mono-and multilingual Dutch language models. *Computational Linguistics in the Netherlands Journal*, 14:525–553.

Ivan Vulić. 2024. Personal communication.

Ivan Vulić, Simon Baker, Edoardo Maria Ponti, Ulla Petti, Ira Leviant, Kelly Wing, Olga Majewska, Eden Bar, Matt Malone, Thierry Poibeau, Roi Reichart, and Anna Korhonen. 2020. Multi-SimLex: A large-scale evaluation of multilingual and crosslingual lexical semantic similarity. *Computational Linguistics*, 46(4):847–897.

Albert Webson and Ellie Pavlick. 2022. Do prompt-based models really understand the meaning of their prompts? In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2300–2344, Seattle, United States. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Kaijie Zhu, Qinlin Zhao, Hao Chen, Jindong Wang, and Xing Xie. 2024. Promptbench: A unified library for evaluation of large language models. *Journal of Machine Learning Research*, 25(254):1–22.

A. Data and Code Availability

The Dutch Multi-SimLex dataset and the Python scripts used for model evaluation are publicly available in the following GitHub repository:

https://github.com/lizzybrans/Dutch-MultiSimLex

B. Model Specifications

Model	Arch.	Params	Corpus	Tokenizer	Scope
Static embeddings					
Word2Vec	Skip-gram	300d	NL Wikipedia (2018)	Gensim	NL
FastText	CBOW+subword	300d	CC + Wiki (NL)	char n-grams	NL
GloVe	Matrix fact.	300d	OSCAR-NL (100k)	Stanford	NL
Encoder models					
BERTje	BERT-base	110M	SoNaR + COW + Wiki	WordPiece (30k)	NL
RobBERT v2	RoBERTa-base	117M	OSCAR-NL (6.6B)	BPE (50k)	NL
RobBERT-2023	RoBERTa-large	355M	OSCAR-NL (2022)	RoBERTa	NL
mBERT	BERT-base	178M	Wikipedia (104L)	WordPiece (110k)	Multi
XLM-R	RoBERTa-base	278M	CC100 (100L)	SentencePiece	Multi
EuroBERT-210M	Llama-like	210M	EU Web/Books	BPE (128k)	EU Multi
EuroBERT-2.1B	Llama-like	2.1B	EU Web/Books	BPE (128k)	EU Multi
Encoder–decoder models					
mT5	T5-base	580M	mC4 (101L)	SentencePiece	Multi
FLAN-T5	T5-base	248M	mC4 + FLAN	SentencePiece	Multi
Decoder-only models					
BLOOM	GPT-like	560M	ROOTS (46L)	byte BPE	Multi
Falcon-7B	Decoder	7B	RefinedWeb	BPE (65k)	EN
Schaapje-2B	Granite	2B	NL Wiki + mC4	Granite tok.	NL

Table 8: Technical specifications of models used in embedding-based evaluation.

C. Layer-wise results

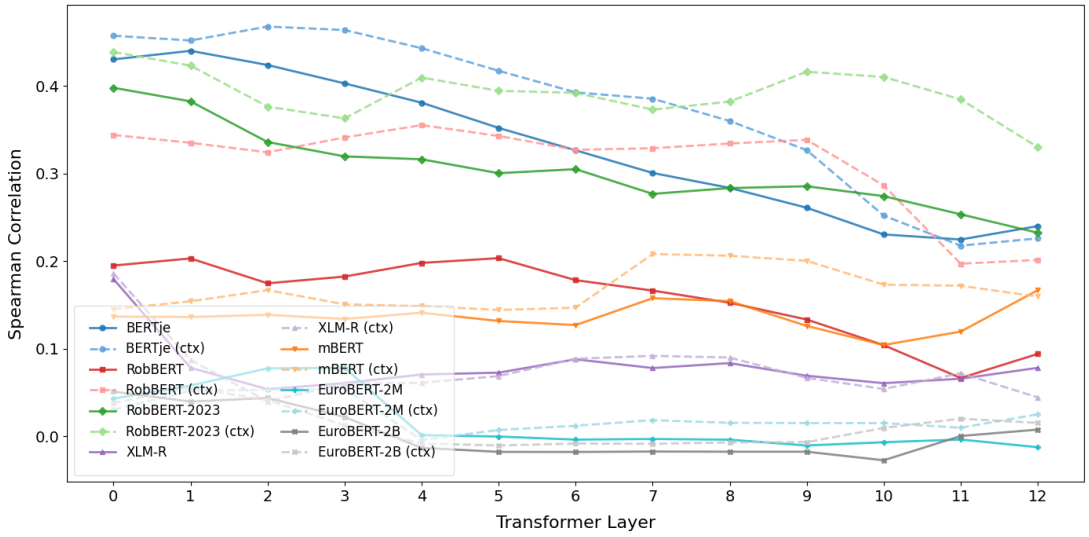


Figure 1: Layer-wise Spearman correlation (ρ) for monolingual and multilingual encoder models on Dutch Multi–SimLex.

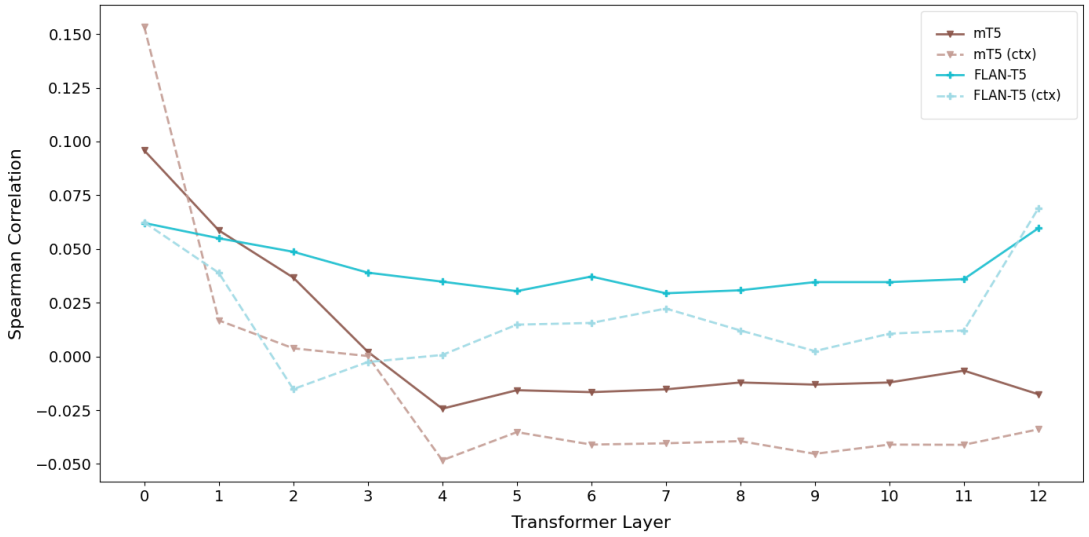


Figure 2: Layer-wise Spearman correlation (ρ) for encoder–decoder models on Dutch Multi–SimLex.

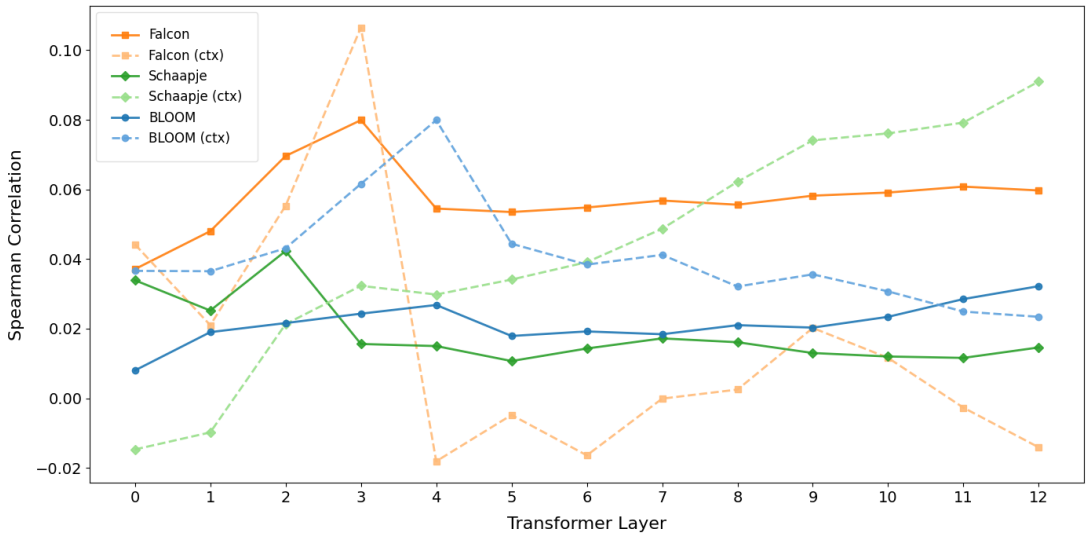
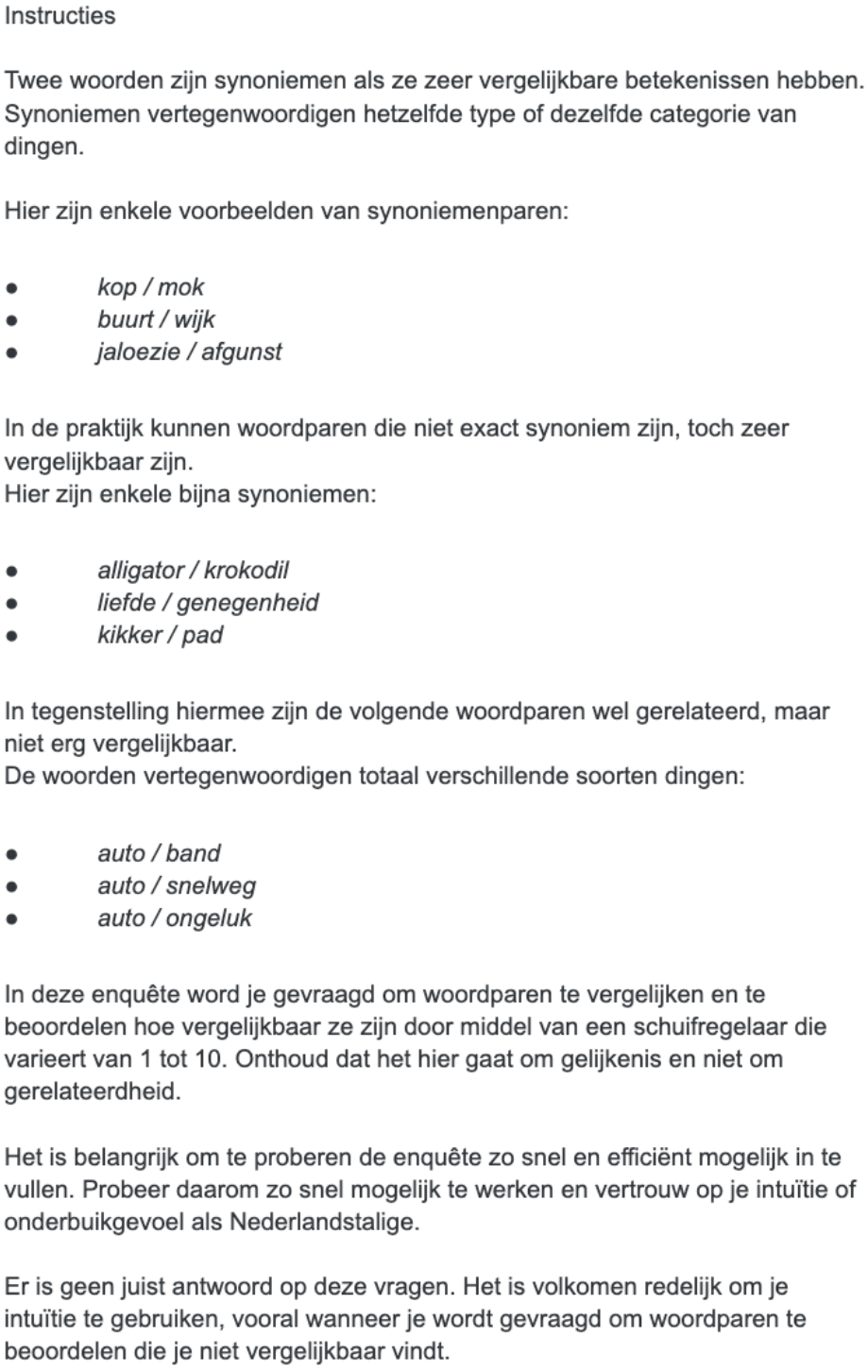


Figure 3: Layer-wise Spearman correlation (ρ) for decoder-only models (Falcon, BLOOM, Schaapje) on Dutch Multi–SimLex.

D. Annotation Details

The instructions provided to the human annotators were translated to Dutch from the original SimLex-999 study (Hill et al., 2015; Brans and Bloem, 2024). As shown in Figure 4, the instructions emphasised that participants should rate word pairs based on semantic similarity rather than topical relatedness and provided clear examples to guide them.



Instructies

Twee woorden zijn synoniemen als ze zeer vergelijkbare betekenissen hebben. Synoniemen vertegenwoordigen hetzelfde type of dezelfde categorie van dingen.

Hier zijn enkele voorbeelden van synoniemenparen:

- *kop / mok*
- *buurt / wijk*
- *jaloezie / afgunst*

In de praktijk kunnen woordparen die niet exact synoniem zijn, toch zeer vergelijkbaar zijn.
Hier zijn enkele bijna synoniemen:

- *alligator / krokodil*
- *liefde / genegenheid*
- *kikker / pad*

In tegenstelling hiermee zijn de volgende woordparen wel gerelateerd, maar niet erg vergelijkbaar.
De woorden vertegenwoordigen totaal verschillende soorten dingen:

- *auto / band*
- *auto / snelweg*
- *auto / ongeluk*

In deze enquête word je gevraagd om woordparen te vergelijken en te beoordelen hoe vergelijkbaar ze zijn door middel van een schuifregelaar die varieert van 1 tot 10. Onthoud dat het hier gaat om gelijkenis en niet om gerelateerdheid.

Het is belangrijk om te proberen de enquête zo snel en efficiënt mogelijk in te vullen. Probeer daarom zo snel mogelijk te werken en vertrouw op je intuïtie of onderbuikgevoel als Nederlandstalige.

Er is geen juist antwoord op deze vragen. Het is volkomen redelijk om je intuïtie te gebruiken, vooral wanneer je wordt gevraagd om woordparen te beoordelen die je niet vergelijkbaar vindt.

Figure 4: Screenshot of the instructions provided to Dutch annotators during the similarity rating task.