

Analyzing and Improving the Efficiency and Interpretability of Neural Language Models

PhD Thesis
Tamás Ficsor

Supervisor: Gábor Berend, PhD
Doctoral School of Computer Science

Department of Artificial Intelligence

Faculty of Science and Informatics

University of Szeged



Szeged
2026

1 Introduction

Modern Natural Language Processing (NLP) has been transformed by large-scale self-supervised neural language models. Transformer-based architectures achieve strong empirical performance across a wide range of language understanding and generation tasks, but their increasing scale also creates important practical and scientific challenges. Large models are costly to train and deploy, their internal computations are difficult to interpret, and their predictions are often overconfident even when they are wrong.

The PhD thesis investigates how these limitations can be addressed by analyzing and restructuring the internal representations of neural language models. Its central premise is that linguistic knowledge is not distributed uniformly across parameters, layers, and representation dimensions. Some components carry disproportionately important information, while others are redundant, entangled, or associated with prediction failures. Exploiting this structure makes it possible to build language models that are more efficient, more interpretable, and more reliable.

The dissertation consists of three major parts. The first part studies efficiency-oriented alternatives to large-scale language models, including low-parameter Transformer models for Hungarian and adaptive layer-wise pruning as a general task driven approach. The second part develops methods for transforming static and contextual representations into explicit semantic spaces aligned with human-interpretable categories. Furthermore, its demonstrating the interpretable nature of sparse representations. The third part extends sparsity-based representation analysis to uncertainty estimation, showing that sparse dictionary atoms (obtained from dense contextual representations) can reveal the sources of model unreliability.

The main results of the dissertation are organized into the following three thesis groups:

- **Efficient language modeling:** compact model design and adaptive pruning for reducing computational cost while preserving task performance.
- **Explicit semantic representation:** transformation methods that expose semantic information in static, contextual, dense, and sparse embedding spaces.
- **Sparsity-based uncertainty estimation:** an interpretable uncertainty estimation framework that links sparse representation components to prediction failures.

2 Efficient Language Modeling

Large Transformer models have become the dominant architecture in NLP, but their computational and memory requirements limit their use in low-resource languages, embedded environments, and practical deployment scenarios. The first thesis group addresses this problem by proposing alternatives that improve efficiency without relying solely on brute-force scaling.

2.1 HuBERTUSz: Low-Parameter Transformer Models for Hungarian

The thesis introduces HuBERTUSz, a family of low-parameter Transformer language models designed for Hungarian language. Hungarian is a morphologically rich language, and generic multilingual models often allocate capacity inefficiently for such languages. HuBERTUSz therefore focuses on monolingual architectural design and training choices that better match the linguistic characteristics of Hungarian.

The HuBERTUSz model family contains compact Transformer variants with substantially fewer parameters than large multilingual

baselines. The models are evaluated on Hungarian named entity recognition, part-of-speech tagging, natural language understanding, and sentiment-related benchmarks. The experiments show that monolingual design can produce competitive results with much lower computational requirements. The medium-sized HuBERTUSz configuration provides a particularly favorable trade-off between model size and performance.

2.2 Adaptive Greedy Layer-wise Pruning

The thesis also proposes an Adaptive Greedy Layer-wise Pruning (AGLP) framework for Transformer models. The method builds on the observation that layers in pre-trained language models are not equally important for every downstream task. Some layers can be removed with little or no loss in accuracy, especially when the remaining model is retrained or repurposed after pruning.

AGLP iteratively identifies layers whose removal causes the smallest decrease in validation performance. After a layer is removed, the model is further trained so that the remaining layers can adapt to the new architecture. This adaptive process makes pruning task-aware and avoids the limitations of static removal strategies.

The evaluation covers both sequence classification (GLUE) and token classification tasks (CoNLL 2003: POS, NER). The results show that substantial layer reduction is possible while preserving a high proportion of the original model performance. The approach provides a practical way to control the trade-off between computational efficiency and downstream accuracy.

3 Transforming Neural Language Representations to Explicit Semantics

The second thesis group focuses on interpretability. Neural language models represent linguistic information in high-dimensional vector spaces, but the semantic content of these spaces is usually latent and distributed across many dimensions. The thesis develops methods for transforming such spaces into explicit semantic bases that are easier to inspect and analyze.

3.1 Semantic Category Construction and Distributional Separation

The interpretability framework starts from semantic categories that approximate human-interpretable concepts. For each category, the method compares the distribution of words belonging to the category with a complementary distribution of other words. The separation between these distributions is measured with statistical distances, most importantly the Hellinger distance, and the resulting scores define semantic directions in the embedding space.

This procedure produces a transformation matrix that reorients an original embedding space toward explicit semantic dimensions. Instead of relying on opaque coordinates, words and contextual tokens can be represented in a basis where dimensions correspond more directly to semantic categories. The method is distribution-agnostic and can be applied across different embedding models and languages.

3.2 Static and Multilingual Embedding Spaces

The thesis applies the semantic transformation approach to static English and Hungarian word embeddings. The results show that

the transformed spaces retain semantic capacity while improving interpretability. In Hungarian, the work evaluates several static embedding spaces and demonstrates that semantic information can be recovered even in morphologically rich settings.

The thesis also introduces a multilingual semantic category generation procedure. This makes it possible to transfer the interpretability framework beyond English and examine how semantic categories are represented across languages. The resulting analyses show that explicit semantic transformations can provide useful tools for comparing embedding spaces and diagnosing their linguistic structure.

3.3 Contextual and Sparse Representations

The framework is further extended to contextual representations produced by neural language models such as BERT-like architectures. Because contextual embeddings vary with the surrounding sentence, their semantic structure is more complex than that of static embeddings. The thesis evaluates how semantic information changes across layers, model sizes, and representation types.

A central result is that sparse representations amplify semantic information. By applying dictionary learning for sparse coding, dense contextual vectors are decomposed into sparse combinations of dictionary atoms. These sparse components tend to localize semantic information more clearly, which improves interpretability while preserving useful representational capacity. This result connects representation sparsity to human-aligned semantic analysis and motivates the use of sparse representations in the third thesis group.

4 Uncertainty Estimation through Sparsity-Induced Methods

The third thesis group addresses reliability. Modern language models can produce confident predictions even when those predictions are incorrect. This is especially problematic in applications where incorrect outputs should be detected, abstained from, or routed to human review. The thesis treats uncertainty not only as an output-level probability, but also as a structural property of learned representations.

4.1 Sparsity-Based Uncertainty Estimation

The thesis introduces Sparsity-Based Uncertainty Estimation (SUE), a framework that estimates uncertainty from sparse decompositions of neural representations. Given hidden states from a trained model, dictionary learning is used to obtain sparse codes. These codes identify which dictionary atoms are active for each sample.

SUE then analyzes the relationship between active atoms and prediction correctness. If certain atoms frequently occur in misclassified samples, they can be interpreted as indicators of risk or uncertainty. Conversely, atoms that frequently occur in correctly classified samples can indicate reliability. This atom-level analysis makes the uncertainty estimate interpretable, because the final score can be traced back to individual sparse components.

4.2 PMI-Based Scoring and Interpretability

The method quantifies the association between dictionary atoms and correctness using pointwise mutual information. For each atom, SUE estimates how strongly it co-occurs with correct and incorrect predictions on a calibration set. The uncertainty score for a new example is computed from the active sparse atoms and their learned associations.

This design has two advantages. First, it provides an uncertainty score that does not require repeated stochastic inference, making it computationally efficient at test time. Second, it connects uncertainty to interpretable representation components. The thesis demonstrates that individual atoms can be inspected to understand why a prediction is considered reliable or uncertain.

4.3 Empirical Evaluation

SUE is evaluated on a range of natural language understanding and sentiment analysis tasks, including adversarial and hard-to-classify settings. It is compared against confidence-based, entropy-based, distance-based, and Bayesian-style uncertainty estimation baselines. The results show that SUE is competitive with, and often superior to, these alternatives according to selective classification metrics such as excess Area Under the Risk-Coverage Curve.

The empirical analysis also demonstrates that the method is relatively robust to hyperparameter choices and efficient after the dictionary has been learned. These findings show that sparse representation structure can support both uncertainty estimation and interpretability, bridging two major themes of the dissertation.

5 Contributions of the Thesis

In the **first thesis group**, my contributions are related to efficient language modeling and model reduction. Detailed discussion can be found in Chapter 3 of the PhD thesis, and the main contributions can be summarized as:

- I/1. The design and evaluation of HuBERTUSz, a family of low-parameter Transformer models for Hungarian language processing.

- I/2. Through comparison of existing mono- and multilingual transformer models on Hungarian benchmarks
- I/3. A greedy and adaptive layer-wise pruning methodology for Transformer models that removes redundant layers without substantial loss in accuracy.
- I/4. Extensive evaluation of model reduction with respect to performance on downstream tasks.

In the **second thesis group**, my contributions are related to interpretable semantic representations. Detailed discussion can be found in Chapter 4 of the PhD thesis. The main contributions are as follows:

- II/1. An improvement of a semantic transformation framework that maps dense, sparse, static and contextual embeddings into explicit, human-aligned semantic spaces with the application of Hellinger distance and kernel density estimation.
- II/2. Introduction of multilingual semantic dataset generation.
- II/3. The evaluation of Hungarian static embeddings and extension to contextual neural networks.
- II/4. The demonstration that sparse representations significantly enhance semantic localization and interpretability.
- II/5. Extensive evaluation of interpretable representations across languages, model sizes, and embedding types.

In the **third thesis group**, my contributions are related to uncertainty estimation and model reliability. Detailed discussion can be found in Chapter 5 of the PhD thesis. Furthermore, the key contributions can be summarized as follows:

- III/1. The introduction of SUE, a novel framework grounded in sparse dictionary learning.

- III/2. A pointwise mutual information-based analysis linking sparse representation components to prediction failures.
- III/3. An interpretable uncertainty scoring mechanism that enables atom-level inspection of uncertainty sources.
- III/4. A large-scale empirical evaluation demonstrating robust uncertainty estimation across commonly used natural language understanding tasks and adversarial settings.

Table 1 summarizes the relation between the thesis points and the corresponding publications.

Table 1: *Correspondence between the thesis points and the author’s publications.*

Publication	Thesis point												
	I/1	I/2	I/3	I/4	II/1	II/2	II/3	II/4	II/5	III/1	III/2	III/3	III/4
[1]	•	•											
[2]			•	•									
[3]					•				•				
[4]						•	•		•				
[5]					•		•	•	•				
[6]										•	•	•	•

The Author's Publications on the Subjects of the Thesis

Full papers in conference proceedings

- [1] **Tamás Ficsor** and Gábor Berend. HuBERTUSz: Alacsony paraméterszámú transzformer modellek létrehozása és kiértékelése magyar nyelvre. In: *XIX. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY2023)*, pp. 417–432, Szegedi Tudományegyetem, Informatikai Intézet, Szeged (2023), ISBN 978-963-306-912-7
- [2] **Tamás Ficsor** and Gábor Berend. Adaptive Greedy Layer Pruning: Iterative Layer Pruning with Subsequent Model Repurposing. In *International Conference on Applications of Natural Language to Information Systems*, pp. 271–284, Springer, 2024.
- [3] **Tamás Ficsor** and Gábor Berend. Interpreting Word Embeddings Using a Distribution Agnostic Approach Employing Hellinger Distance. In *Text, Speech, and Dialogue*, pp. 197–205, Springer International Publishing, 2020.
- [4] **Tamás Ficsor** and Gábor Berend. Analysing the Semantic Content of Static Hungarian Embedding Spaces. In: *XVII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY2021)*, pp. 91–105, Szegedi Tudományegyetem, Informatikai Intézet, Szeged (2021), ISBN 978-963-306-781-9
- [5] **Tamás Ficsor** and Gábor Berend. Changing the Basis of Contextual Representations with Explicit Semantics. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Student Research Workshop*, pp. 235–247,

Association for Computational Linguistics, 2021.

- [6] **Tamás Ficsor** and Gábor Berend. SUE: Sparsity-Based Uncertainty Estimation via Sparse Dictionary Learning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 32912–32929, Association for Computational Linguistics, 2025.

Other References

- [1] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017.
- [2] Devlin, J., Chang, M.W., Lee, K., and Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, 2019.
- [3] Mairal, J., Bach, F., Ponce, J., and Sapiro, G. Online dictionary learning for sparse coding. In *Proceedings of the 26th Annual International Conference on Machine Learning*, 2009.

6. Összefoglalás

A doktori értekezésemben azt vizsgálom, hogy a modern neurális nyelvi modellek belső reprezentációinak elemzésével és átszervezésével azok miként tehetők hatékonyabbá, értelmezhetőbbé és megbízhatóbbá. Bár a korszerű nyelvi modellek erős empirikus teljesítményt érnek el, gyakran jelentős számítási erőforrást igényelnek, nehezen értelmezhetők, és hibás predikciók esetén is túlzottan magabiztosak lehetnek. A dolgozat e problémákat reprezentációtanulási, ritkaságot előidéző és információelméleti módszerek összekapcsolásával kezeli.

A tézis három fő témakörből áll. Az első része a hatékony nyelvi modellezésre összpontosít. Bemutatja a HuBERTUSz modellecsaládot, amely alacsony paraméterszámú, magyar nyelvre optimalizált Transformer modellekből áll, valamint egy mohó-adaptív réteg szintű redukciós eljárást, amely a redundáns rétegek eltávolításával csökkenti a modellek méretét, a teljesítmény jelentős romlása nélkül.

A második rész a neurális nyelvi reprezentációk értelmezhetőségével foglalkozik. A dolgozat olyan transzformációs keretrendszert vezet be, amely statikus és kontextuális beágyazásokat explicit, ember által értelmezhető szemantikai terekbe képez le. A módszer szemantikai és komplementer eloszlások elkülönülését méri, majd ezen távolságok alapján hoz létre új szemantikai bázisokat. Az eredmények azt mutatják, hogy a ritka reprezentációk felerősítik a szemantikai lokalizációt és javítják az értelmezhetőséget.

A harmadik rész a modellmegbízhatóságot vizsgálja. A dolgozat bevezeti a Sparsity-Based Uncertainty Estimation (SUE) módszert, amely ritka szótártanulással bontja fel a neurális reprezentációkat, majd a szótáratomok és a predikciós hibák kapcsolatát pontonkénti kölcsönös információval elemzi. A módszer hatékony és értelmezhető bizonytalansági pontszámot ad, és több nyelvértési, szentimentelemzési, valamint ellenséges beállításban is versenyképes vagy jobb eredményt ér el a hagyományos bizonytalanságbecslési eljárásokhoz képest.

Nyilatkozat

Ficsor Tamás “Analyzing and Improving the Efficiency and Interpretability of Neural Language Models” című PhD disszertációjában a következő eredményekben Ficsor Tamás hozzájárulása volt a meghatározó:

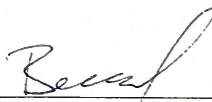
- I/1. A HuBERTUSz modellcsalád (amely alacsony paraméterszámú, kifejezetten magyar nyelvfeldolgozási feladatokra) előtanításának megtervezése és kivitelezése.
- I/2. Összehasonlító elemzés elvégezése meglévő egynyelvű és többnyelvű Transformer modelleken, magyar nyelvű benchmark adathalmazok és feladatok felhasználásával.
- I/3. Egy adaptív, mohó rétegszintű modellredukciós módszer kidolgozása Transformer alapú modellekhez, amely a redundáns rétegek eltávolításával csökkenti a modell méretét.
- I/4. Átfogó kiértékelés a modellredukció hatásának különböző downstream feladatokon. Különös tekintettel a modellméret, a számítási igény és a feladatspecifikus teljesítmény közötti kompromisszumokra.
- II/1. Egy szemantikai transzformációs keretrendszer továbbfejlesztése, amely sűrű, ritka, statikus és kontextuális beágyazásokat explicit, szemantikai terekbe képez le Hellinger-távolság és kernelsűrűségbecslés alkalmazásával.
- II/2. Egy többnyelvű szemantikai adathalmaz-előállítási eljárásának kidolgozása, amely lehetővé teszi az explicit szemantikai reprezentációk különböző nyelveken történő vizsgálatát.
- II/3. Magyar statikus beágyazási terek kiértékelése, és a szemantikai transzformációs módszer kiterjesztése kontextuális neurális nyelvi reprezentációkra.
- II/4. A ritka reprezentációkkal kapcsolatosan feltevések igazolása, miszerint javítják a szemantikai információ lokalizációját, és ezáltal növelik a neurális reprezentációk értelmezhetőségét.

- II/5. Átfogó kiértékelés elvégzése különböző nyelveken, modellméreteken, neurális hálózati rétegeken és beágyazástípusokon az explicit szemantikai reprezentációk összehasonlítása érdekében.
- III/1. A SUE eljárása kidolgozása, amely ritka szótártanuláson alapuló bizonytalanságbecslési keretrendszert határoz meg.
- III/2. Pontonkénti kölcsönös információn alapuló elemzési módszer alkalmazása, amely összekapcsolja a ritka reprezentációk egyes komponenseit a helyes és hibás modellpredikciókkal.
- III/3. A módszerrel járó értelmezhetőségi folyamat vizsgálata, amely lehetővé teszi a bizonytalanság forrásainak szótáratom-szintű vizsgálatát és azok valós példányhoz történő visszavezetését.
- III/4. Empirikus kiértékelés elvégzése általánosan alkalmazott természetesnyelv-megértési, szentimentelemzési és ellenséges feladatokon. Ezek az eredmények Ficsor Tamás PhD disszertációján kívül más tudományos fokozat megszerzésére nem használhatók fel.

Dátum: 2026.07.06



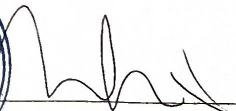
Ficsor Tamás
doktorjelölt



Berend Gábor
témavezető

Az Informatika Doktori Iskola vezetője kijelenti, hogy jelen nyilatkozatot minden társszerzőhöz eljuttatta, és azzal szemben egyetlen társszerző sem emelt kifogást.

Dátum: 2026.07.06.


Doktori Iskola vezetője

