Budapest University of Technology and Economics
Faculty of Electrical Engineering and Informatics
Department of Automation and Applied Informatics

Morphology in the Age of Pre-trained Language Models

Ph.D. Thesis Booklet

Judit Ács

Thesis supervisor:
András Kornai, D.Sc.

Budapest
2025

Introduction

The field of natural language processing (NLP) has adopted deep learning methods in the past 15 years. Nowadays the state-of-the-art in most NLP tasks is some kind of neural model, often the fine-tuned version of a pre-trained language model. The efficacy of these models is demonstrated on various English benchmarks and increasingly, other monolingual and multilingual benchmarks. In this dissertation I explore the application of deep learning models on low level tasks, particularly morphosyntactic[1] tasks in multiple languages.

The first part of this dissertation (Theses 1 and 2) explores the application of deep learning models for classical morphosyntactic tasks such as morphological inflection and generation in dozens of languages with special focus on Hungarian.

The second part of this dissertation (Theses 3 to 5) deals with pre-trained language models, mostly models from the BERT (Devlin et al., 2019) family. I include some experiments on GPT-4o and GPT-4o-mini. These models show excellent performance on various tasks in English and some high density languages. However, their evaluation in medium and low density languages is lacking. I present a methodology for generating morphosyntactic benchmarks in arbitrary languages and I analyze multiple BERT-like models in detail. My main tool for analysis is the probing methodology (Belinkov, 2021).

Research methodology

The main research question of this dissertation is how well deep learning models handle morphology in many languages. I examine this question from two aspects. The first (Part I., Theses 1 and 2) one looks at small deep learning models trained specifically for low level tasks. The second one (Part II., Theses 3 to 5) uses mid-sized pre-trained language models as encoders and I examine their morphosyntactic content.

[1] Morphology is the study of morphemes, the smallest meaningful units of language, while syntax is the study of sentence formation. Morphosyntax refers to the fact that morphemes often have sentence-level functions and the two fields of linguistics are far from independent.

Part I. of my dissertation uses end-to-end training for hand assembled deep learning models. The models are trained from scratch on small to medium amounts of annotated data. I applied standard training procedures except when noted otherwise.

Probing is my main method of research in Part II. Probing, as a tool for model inspection, has been applied to many applications including morphology but not as extensively as I do in my dissertation. Probing is a simple and intuitive way of asserting model contents but critics have pointed out some flaws (Belinkov, 2021). As part of my ablations, I showed that most of these problems have limited effect on my studies and my results are independent of the type of probing and its parameters.

I introduce novel sentence perturbations which remove some information from the sentence and by retraining the probes, I can determine whether this information source was necessary. I then extend this analysis with Shapley values borrowed from game theory.

My contributions are summarized in the following 5 theses.

Thesis 1

I demonstrated that encoder-decoder (a.k.a. sequence-to-sequence or seq2seq) models are well-suited for morphological inflection and generation. This holds for type-level and sentence-level tasks in multiple languages.

Subtheses:

1.1 I collected and prepared a silver standard Hungarian dataset for morphological inflection and analysis using a high quality rule-based analyzer.

1.2 I implemented two types of sequence-to-sequence or encoder-decoder models with attention: LSTM with soft attention and LSTM with hard attention.

1.3 I trained and evaluated the models on the Hungarian dataset.

1.4 I developed two new types of encoder-decoder models for morphological inflection. The first model is a double encoder-single decoder model for type-level inflection. The second model is complex multi-encoder model for sentence-level inflection. I participated in the CoNLL-SIGMORPHON 2018 Shared Task with these models as an individual team. Task 1 was type-level inflection in over 100 languages, Task 2 was inflection in context (sentence) in 7 languages. I placed 3rd and 2nd in the two tasks respectively with multi-encoder and single-decoder seq2seq neural networks.

These contributions were published in Ács (2018) and they are my sole contribution.

Thesis 2

I adapted a differentiable weighted finite state RNN to sequence-to-sequence tasks and showed that neural pattern matching can extract morphosyntactic patterns in multiple languages when used as an encoder for morphological inflection and analysis.

Subtheses:

2.1 I reimplemented the Soft Patterns or SoPa neural model (Schwartz et al., 2018), a restricted and differentiable finite state automaton model originally used for text classification tasks. I added an LSTM decoder and used SoPa as an encoder-decoder model. This model uses pattern matching on the encoder side.

2.2 I applied the model to type-level morphological analysis and inflection in 12 typologically diverse languages.

2.3 I extracted patterns (character sequences) from trained SoPa models and manually examined their linguistic plausibility.

2.4 I introduced a model similarity metric defined for a pair of SoPa models. This metric allows comparing different tasks that use the same input

data. The higher the similarity between two tasks, the more likely they are to rely on the same patterns.

These contributions were published in Ács and Kornai (2020). The paper was awarded the best paper award at the Hungarian Computational Linguistics Conference in 2020. The implementation is entirely my contribution.

Thesis 3

I developed a new methodology for morphosyntactic probing. It relies on CoNLL-U formatted data which is widely available in the Universal Dependencies Treebanks, therefore my methodology is applicable to a large number of languages and morphosyntactic tags. Using my probing methodology I showed that pretrained language models (PLMs) trained on unannotated text learn morphology. PLMs' representations retain morphosyntactic information across a large set of typologically diverse languages and multiple tasks.

Subtheses:

3.1 I introduced the largest multilingual morphosyntactic probing dataset with 247 tasks in 42 families from 10 language families. I define a probing sample as a triplet of a sentence, a particular token called *target token* in the sentence and a morphosyntactic tag corresponding to the target token. I used the Universal Dependencies Treebank (Nivre et al., 2018) to generate probing samples according to this definition.

3.2 I evaluated 6 multilingual PLMs and analyze two, mBERT and XLM-RoBERTa in detail. I compared them to various baselines and show that PLMs indeed learn morphosyntactic information.

3.3 I evaluated GPT-4o and GPT-4o-mini on the test set of the probing dataset via prompting.

3.4 I examined the tokenizer of PLMs and showed that although multilingual models support over 100 languages, the tokenizer works better on languages that use the Latin script, particularly English. (Ács, 2019)

3.5 Probing as an analysis tool for blackbox models has been criticized for various reasons (Belinkov, 2021). I introduced more than 10 ablation methods and showed that the conclusions drawn from morphosyntactic probing are robust.

3.6 The token-level usage of PLMs requires a way of handling tokens split into multiple subwords represented by multiple vectors. A pooling function (parametric on non-parametric) may be used to infer a single vectors by token. I showed that the choice of pooling function matters, especially for feature extraction (when the PLM is not fine-tuned). I compared 9 pooling functions in 7 languages and 3 tasks (Ács et al., 2021a).

These contributions were published in (Ács, 2019; Ács et al., 2021a; Ács et al., 2023). The experiment design was done in collaboration with my coauthors and the implementation is my sole contribution.

Thesis 4

I used my morphosynactic probing dataset and methodology to demonstrate that monolingual PLMs are better in their respective languages than multilingual PLMs but the difference is small and often not statistically significant. Moreover both monolingual and multilingual PLMs can be successfully transferred to new languages as long as the new language uses the same writing system.

Subtheses:

4.1 I analyzed 5 PLMs with Hungarian support and I found that HuBERT (Nemeskey, 2021), a Hungarian-only model is better at morphological, POS and NER tagging than the multilingual models, especially distilBERT, but the margin is small. (Ács et al., 2021)

4.2 I extended this study to the languages of the Uralic family (Ács et al., 2021b). I drew similar conclusions in Estonian and Finnish, the only Uralic languages with dedicated monolingual PLMs, as in Hungarian.

4.3 I trained POS and NER tagging models in minority Uralic languages by transferring multilingual and monolingual models. The cross-language models are surprisingly successful despite the extremely small training data available in some languages. The new models appear to be state-of-the-art without any language-specific effort.

These contributions were published in (Ács et al., 2021; Ács et al., 2021b). The experiment design and the result analysis was done in collaboration with my coauthors. The implementation is my contribution.

Thesis 5

I refined my probing analysis with perturbations that aim to find the exact location of morphosyntactic information in a sentence. The systematic removal of certain information (perturbations) reveals where the information is stored. I used Shapley values to quantify the role of context in morphosyntax and the results often agree with linguistic intuitions.

Subtheses:

5.1 I introduced a set of perturbation methods that remove some source of information from a sentence. I retrained the morphosyntactic probes (cf. Thesis 3) on the 247 probing tasks. The results offer insight on where the information is stored in the sentence. We can often find linguistic explanations for them.

5.2 I used the results of perturbations as features for clustering the languages. Such clustering tends to group languages from the same family in the same cluster with some notable exceptions. It also tends to cluster typologically similar but unrelated languages in the same cluster.

5.3 I defined a sentence as a 9-player coalition game where the players are tokens or groups of tokens relative to the target token in the morphosyntactic probe. I applied the Shapley framework on each probing task.

5.4 I analyzed the Shapley values from mBERT, XLM-RoBERTa and an LSTM baseline and found that the inter-model correlation is high which demonstrates that the Shapley values are more descriptive of linguistic structure than of the models. I identified the outlier tasks and I found that the Shapley values often confirm the linguistic properties of the particular language and task.

These contributions were published in (Ács et al., 2023). The experiment design and the result analysis was done in collaboration with my coauthors. The implementation is my contribution.

Challenges and limitations

Part I. of this dissertation dates back to the time of smaller task-specific neural models which are less widely used nowadays especially for building new applications. The limitations at the time were mainly the availability of annotated training data or a high quality rule-based morphological analyzer suitable for data generation.

The main limitations of Part II. pertain to the nature of the tasks I examined. Morphosyntactic tasks are low-level tasks and the performance of certain models on these tasks may not give a lot of insight about what we can expect on high level tasks which are generally accepted as benchmarks when available. Unfortunately the large majority of languages, even ones with sufficient amounts of raw text data for LLM training, lack such benchmarks and I had to work with what was available.

The second important limitation of Part II. is data quality. One source of data errors is the Universal Dependencies Treebank, my main source, itself. Although generally a high quality resource, it is composed of diverse treebanks, even within the same language, and it is bound to have errors and inconsistencies. The high number of languages I studied made it impossible to do specialized quality control. The second source of errors is the result of the sampling method itself.

The third major limitation of Part II. is the lack of fine-tuning aside from a small ablation study (Subthesis 3.5) and the evaluation of Uralic languages

(Thesis 4). The main reason for this is efficiency. Fine-tuning would have resulted in an 80-fold increase in training time and an even larger increase in evaluation time due to the necessity of reloading the fine-tuned BERT models each time.

Practical applications

The small deep learning models I explore in Thesis 1 have good to excellent performance on morphological tasks with limited training data. The vast majority of the world's languages have minimal to no training data and creating new annotated datasets is costly or often impossible due to the lack of linguistic experts. Encoder-decoder models such as the ones I examine, are well-suited for bootstrapping dataset creation.

SoPa, the subject of Thesis 2, had a great potential as an interpretable model with simple reasoning capabilities but it was greatly overshadowed by Transformer models and it remains an interesting proof-of-concept with very few practical applications.

Part II. of my dissertation is a large scale evaluation of multilingual and monolingual models which in itself is a practical application. When we choose a model from multiple options, one of the most important factors is its performance in early evaluations. My morphosyntactic probing methodology is applicable to any model including generative models but it is better suited for encoder and encoder-decoder models. It is also applicable to any language as long as some morphologically annotated data is available.

Thesis 4 demonstrates two sets of experiments that directly target one language or a language family. This type of comparison is applicable to other scenarios such as other languages and domains. I also explore unsupported languages with extremely limited amounts of annotated data and demonstrate that that models can be transferred with good performance. The transferred models are usable as POS and NER taggers, often the only options available in low-resource languages.

Thesis 5 explores the applicability of Shapley values for fine-grained analysis. The output of this analysis is elegant and interpretable but the computational complexity is a serious limitation for wide scale use.

Publications related to this dissertation

Judit Ács. 2018. BME-HAS system for CoNLL–SIGMORPHON 2018 shared task: Universal morphological reinflection. *Proceedings of the CoNLL SIGMORPHON 2018 Shared Task: Universal Morphological Reinflection*, pages 121–126.

Judit Ács. 2019. Exploring BERT's vocabulary. http://juditacs.github.io/2019/02/19/bert-tokenization-stats.html. Accessed: 2021-05-14.

Judit Ács, Endre Hamerlik, Roy Schwartz, Noah A. Smith, and Andras Kornai. 2023. Morphosyntactic probing of multilingual BERT models. *Natural Language Engineering*, page 1–40.

Judit Ács, Ákos Kádár, and Andras Kornai. 2021a. Subword pooling makes a difference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2284–2295, Online. Association for Computational Linguistics.

Judit Ács and András Kornai. 2020. The role of interpretable patterns in deep learning for morphology. In *XVI. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY2020)*, pages 171–179, Szeged.

Judit Ács, Dániel Lévai, and Andras Kornai. 2021b. Evaluating transferability of BERT models on Uralic languages. In *Proceedings of the Seventh International Workshop on Computational Linguistics of Uralic Languages*, pages 8–17, Syktyvkar, Russia (Online). Association for Computational Linguistics.

Judit Ács, Dániel Lévai, Dávid Márk Nemeskey, and András Kornai. 2021. Evaluating contextualized language models for Hungarian. In *XVII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY2020)*, Szeged.

Ádám Kovács, Judit Ács, András Kornai, and Gábor Recski. 2020. Better together: Modern methods plus traditional thinking in NP alignment. In *Proc. LREC 2020*, pages 3635—-3639.

Ádám Kovács, Evelin Ács, Judit Ács, András Kornai, and Gábor Recski. 2019. BME-UW at SRST-2019: Surface realization with interpreted regular tree grammars. In *Proceedings of the 2nd Workshop on Multilingual Surface Realisation (MSR 2019)*, pages 35–40, Hong Kong, China. Association for Computational Linguistics.

Tiago Pimentel, Maria Ryskina, Sabrina J. Mielke, Shijie Wu, Eleanor Chodroff, Brian Leonard, Garrett Nicolai, Yustinus Ghanggo Ate, Salam Khalifa, Nizar Habash, Charbel El-Khaissi, Omer Goldman, Michael Gasser, William Lane, Matt Coler, Arturo Oncevay, Jaime Rafael Montoya Samame, Gema Celeste Silva Villegas, Adam Ek, Jean-Philippe Bernardy, Andrey Shcherbakov, Aziyana Bayyr-ool, Karina Sheifer, Sofya Ganieva, Matvey Plugaryov, Elena Klyachko, Ali Salehi, Andrew Krizhanovsky, Natalia Krizhanovsky, Clara Vania, Sardana Ivanova, Aelita Salchak, Christopher Straughn, Zoey Liu, Jonathan North Washington, Duygu Ataman, Witold Kieraś, Marcin Woliński, Totok Suhardijanto, Niklas Stoehr, Zahroh Nuriah, Shyam Ratan, Francis M. Tyers, Edoardo M. Ponti, Grant Aiton, Richard J. Hatcher, Emily Prud'hommeaux, Ritesh Kumar, Mans Hulden, Botond Barta, Dorina Lakatos, Gábor Szolnok, Judit Ács, Mohit Raj, David Yarowsky, Ryan Cotterell, Ben Ambridge, and Ekaterina Vylomova. 2021. SIGMORPHON 2021 shared task on morphological reinflection: Generalization across languages. In *Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 229–259, Online. Association for Computational Linguistics.

Gábor Recski, Ádám Kovács, Kinga Gémes, Judit Ács, and Andras Kornai. 2020. BME-TUW at SR'20: Lexical grammar induction for surface realization. In *Proceedings of the Third Workshop on Multilingual Surface Realisation*, pages 21–29, Barcelona, Spain (Online). Association for Computational Linguistics.

Gábor Szolnok, Botond Barta, Dorina Lakatos, and Judit Ács. 2021. BME submission for SIGMORPHON 2021 shared task 0. A three step training approach with data augmentation for morphological inflection. In *Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics,*

Phonology, and Morphology, pages 268–273, Online. Association for Computational Linguistics.

Other publications

Judit Ács. 2014. Pivot-based multilingual dictionary building using Wiktionary. In *The 9th edition of the Language Resources and Evaluation Conference*.

Judit Ács. 2015. Synonym acquisition from translation graph. In *XI. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2015)*, pages 14–21. Szegedi Tudományegyetem Informatikai Tanszékcsoport.

Judit Ács, Gábor Borbély, Márton Makrai, Dávid Nemeskey, Gábor Recski, and András Kornai. 2018. Hibrid nyelvtechnológiák. *Magyar Tudomány*, 6.

Judit Ács, László Grad-Gyenge, and Thiago Bruno Rodrigues de Rezende Oliveira. 2015. A two-level classifier for discriminating similar languages. In *Proceedings of the Joint Workshop on Language Technology for Closely Related Languages, Varieties and Dialects*, pages 73–77, Hissar, Bulgaria. Association for Computational Linguistics.

Judit Ács and József Halmi. 2016a. Comparing diacritic restoration methods for Hungarian. In *Proceedings of the Automation and Applied Computer Science Workshop 2016 : AACS'16*. Budapest University of Technology and Economics.

Judit Ács and József Halmi. 2016b. Hunaccent: Small footprint diacritic restoration for social media. In *Proceedings of the First Workshop on Normalisation and Analysis of Social Media Texts (NormSoMe)*.

Judit Ács and Ágota Illyés. 2015. Language detection and generation. In *Proceedings of the Automation and Applied Computer Science Workshop 2015 : AACS'15*. Budapest University of Technology and Economics.

Judit Ács and András Kornai. 2016. Evaluating embeddings on dictionary-based similarity. In *Proceedings of the First Workshop on Evaluating Vector-Space Representations for NLP (RepEval)*, pages 78–82.

Judit Ács, Dávid Nemeskey, and András Kornai. 2017. Identification of disaster-implicated named entities. In *Proceedings of the First International*

Workshop on Exploitation of Social Media for Emergency Relief and Preparedness.

Judit Ács, Dávid Márk Nemeskey, and Gábor Recski. 2017. Building word embeddings from dictionary definitions. In Katalin Mády Beáta Gyuris and Gábor Recski, editors, *K + K = 120: Papers dedicated to László Kálmán and András Kornai on the occasion of their 60th birthdays*. Research Institute for Linguistics, Hungarian Academy of Sciences (RIL HAS).

Judit Ács, Katalin Pajkossy, and András Kornai. 2013. Building basic vocabulary across 40 languages. In *Proceedings of the Sixth Workshop on Building and Using Comparable Corpora*, pages 52–58, Sofia, Bulgaria. Association for Computational Linguistics.

Judit Ács and Géza Velkey. 2017. Comparing word segmentation algorithms. In *Proceedings of the Automation and Applied Computer Science Workshop 2017 : AACS'17*. Budapest University of Technology and Economics.

Botond Barta, Endre Hamerlik, Milán Konor Nyist, and Judit Ács. 2025. HuAMR: A Hungarian AMR parser and dataset. In *XXI. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2025)*, pages 185–196. Szegedi Tudományegyetem Informatikai Tanszékcsoport.

Botond Barta, Dorina Lakatos, Attila Nagy, Milán Konor Nyist, and Judit Ács. 2023. HunSum-1: an abstractive summarization dataset for Hungarian. In *XIX. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2023)*, pages 231–243. Szegedi Tudományegyetem Informatikai Tanszékcsoport.

Botond Barta, Dorina Lakatos, Attila Nagy, Milán Konor Nyist, and Judit Ács. 2024. From news to summaries: Building a Hungarian corpus for extractive and abstractive summarization. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7503–7509, Torino, Italia. ELRA and ICCL.

Dániel Huszti and Judit Ács. 2017. Entitásorientált véleménykinyerés magyar nyelven. In *XIII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY*

2017), pages 240–250. Szegedi Tudományegyetem Informatikai Tanszékcsoport.

András Kornai, Judit Ács, Márton Makrai, Dávid Márk Nemeskey, Katalin Pajkossy, and Gábor Recski. 2015. Competence in lexical semantics. In *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics*, pages 165–175, Denver, Colorado. Association for Computational Linguistics.

Attila Nagy, Dorina Lakatos, Botond Barta, and Judit Ács. 2023a. TreeSwap: Data augmentation for machine translation via dependency subtree swapping. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 759–768, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Attila Nagy, Dorina Petra Lakatos, Botond Barta, Patrick Nanys, and Judit Ács. 2023b. Data augmentation for machine translation via dependency subtree swapping. In *XIX. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2023)*. Szegedi Tudományegyetem Informatikai Tanszékcsoport.

Attila Nagy, Patrick Nanys, Konrád Balázs Frey, Bence Bial, and Judit Ács. 2022. Syntax-based data augmentation for Hungarian-English machine translation. In *XVIII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2022)*. Szegedi Tudományegyetem Informatikai Tanszékcsoport.

Gergely Dániel Németh and Judit Ács. 2018. Hyphenation using deep neural networks. In *XIV. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2018)*. Szegedi Tudományegyetem Informatikai Tanszékcsoport.

Gábor Recski and Judit Ács. 2015. MathLingBudapest: Concept networks for semantic similarity. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 138–142, Denver, Colorado. Association for Computational Linguistics.

References

Yonatan Belinkov. 2021. Probing Classifiers: Promises, Shortcomings, and Advances. *arXiv:2102.12452 [cs]*. ArXiv: 2102.12452.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

Dávid Márk Nemeskey. 2021. Introducing huBERT. In *XVII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2021)*, pages 3–14, Szeged.

Joakim Nivre, Mitchell Abrams, Željko Agić, et al. 2018. Universal Dependencies 2.3. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Roy Schwartz, Sam Thomson, and Noah A. Smith. 2018. SoPa: Bridging CNNs, RNNs, and Weighted Finite-State Machines. In *Proc. 56th ACL Annual Meeting*, pages 295–305, Melbourne, Australia.