Annales Mathematicae et Informaticae
58 (2023) pp. 69–80
DOI: 10.33039/ami.2023.08.009
URL: https://ami.uni-eszterhazy.hu

A pseudonymization tool for Hungarian

Péter Hatvani[a b], **László János Laki**[b], **Zijian Győző Yang**[b]

[a]Pázmány Péter Catholic University Faculty of Humanities and Social Sciences,
Doctoral School of Linguistics
hatvani.peter@hallgato.ppke.hu

[b]Hungarian Research Centre for Linguistics
{yang.zijian.gyozo,laki.laszlo}@nytud.hu

Abstract. In today's world, the volume of documents being generated is growing exponentially, making the protection of personal data an increasingly crucial task. Anonymization plays a vital role in various fields, but its implementation can be challenging. While advancements in natural language processing research have resulted in more accurate named entity recognition (NER) models, relying on an NER system to remove names from a text may compromise its fluency and coherence. In this paper, we introduce a novel approach to pseudonymization, specifically tailored for the Hungarian language, which addresses the challenges associated with maintaining text fluency and coherence. Our method employs a pipeline that integrates various NER models, morphological parsing, and generation modules. Instead of merely recognizing and removing named entities, as in conventional approaches, our pipeline utilizes a morphological generator to consistently replace names with alternative names throughout the document. This process ensures the preservation of both text coherence and anonymity. To assess the efficacy of our method, we conducted evaluations on multiple corpora, with results consistently indicating that our pipeline surpasses traditional approaches in performance. Our innovative approach paves the way for new pseudonymization possibilities across a diverse range of fields and applications.

Keywords: Pseudonymization, Named entity recognition (NER), Morphological generation

AMS Subject Classification: 68T50, 68T07

Submitted: July 31, 2023
Accepted: August 7, 2023
Published online: August 19, 2023

1. Introduction

The GDPR (General Data Protection Act) [3] of the European Union enforces stricter than ever rules on handling personal information and information that can be traced back to the subject. Luckily, data can not only be stored in a completely anonymised way but also in a pseudoanonymised way, still, as per the definition of the law:

> 'pseudonymization' means the processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organisational measures to ensure that the personal data are not attributed to an identified or identifiable natural person;

The most important aspect of the pseudoanonymisation is that after the process direct identification can not remain in the text. As for indirect identification - location, title, martial status - these information can remain and will remain in the text in our implementation.

The paper introduces a pseudononymization tool for Hungarian that integrates different named entity recognition, morphological parsing and generation modules. Instead of simply recognizing and removing named entities, the tool replaces found names with other names consistently throughout a given document. The tool uses huSpacy [14] and emMorph [11], a Hungarian morphological analyzer, to ensure that the results are consistent in several testing corpora. The tool is designed to be used for various use cases such as legal, medical documents and other types of sensitive texts. However, the testing of the model has been limited to crawled comments, excerpt from programmes of the Hungarian Kossuth Radio, excerpts from NerKor [17] from the news and the wikipedia parts and Hungarian literature, which serves as a proof of concept for the ability of the tool to maintain coherence and fluency in the anonymized text.

Our pseudononymization pipeline tool is freely available in our github site[1] with Apache 2.0 license.

2. Related works

Anonymization is an important task in many fields, with healthcare being one of the main areas where it is essential. In medical documents, anonymization methods need to be applied to protect patients' privacy [1, 2, 15, 16]. For Hungarian, Kinga Mátyus [7] conducted rule-based anonymization in a sociolinguistic research.

Various approaches have been proposed for anonymizing medical data. In one study [1], an enhanced method utilizing asymmetric encryption was suggested to separate the duties of pseudonymization and de-pseudonymization. This method

[1] `https://github.com/nytud/pseudo-anonimization`

proposed a secure and controlled process that allows authorized third parties (ombudsmen) to de-pseudonymize patients when necessary, thereby bridging the gap between bench and bedside in translational research while preserving patient privacy.

Another study [16] focused on the use of pseudonymization for retrospective research, quality assurance, and education. By replacing all person-related data within a data record with an artificial identifier, pseudonymization allows for the linking of medical data and patient identification data under specific, predefined, and controllable conditions. Consequently, medical data can be shared with third parties without enabling them to identify the individual patients.

A third study [15] introduced a system called PIPE (pseudonymization of information for privacy in e-health), which securely integrates primary and secondary usage of health data, addressing the shortcomings of existing approaches. PIPE can serve as a foundation for implementing secure electronic health record (EHR) architectures or as an extension to current systems, effectively preventing health data misuse while enhancing communication between healthcare providers and, in turn, improving patient care quality and reducing costs.

Anonymisation based on NER model and morphological tools are rare. For Hungarian, HuSpacy [14] is a spaCy library providing industrial-strength Hungarian language processing facilities. The huSpacy pipeline contains a tokenizer, a sentence splitter, a lemmatizer, a morphological tagger, a dependency parser and a named entity recognition module. The morphological analyzer achieved approximately 90% accuracy in token classification tasks, and the NER module achieved 80.75 F1 score on NerKor [17]. From the heaps of morphological analyzers for Hungarian we have chosen to use huSpacy and emMorph (eMagyar[2]). The emMorph [11] is an Hungarian morphological analyzer that uses Humor [10] unification morphology. The advantage of the spaCy tool is that it is fast and it can produce results without GPU acceleration as well as with it. Three different models are present of HuSpacy, two of them are based on the huBERT [9] model. One of the best Hungarian morphological annotation tool is the PurePos 2.0 [13] that uses emMorph morphology analyzer. The PurePos achieved 96.72% accuracy on part-of-speech recognition tasks. Also emBERT [8] is a framework that was used with a multlimodal BERT model to finetune NER and noun phase recognition (NP) models. On NerKor, the emBERT could achieve 92.09 F1 score, due to time constraints emBERT fell out of this round of comparison, but it showed promise being a freely available NER model that finetuned by [22] on NerKor and gained a 90.18 F1 score.

There are some morphological analyzers for Hungarian that can be used for generation as well. HunSpell can be used for this task in two ways: it generates word forms by typing the lemma and the features, or typing the lemma and an example word. Hunmorph [19] and Morphdb.hu [20] are also suitable for morphological generation. Hunmorph-foma[3] uses the morphological tagset of HunMorph and it is

[2] https://e-magyar.hu

[3] https://github.com/r0ller/hunmorph-foma

based on the foma generator [4]. The main problem of these tools that they are not freely available or use a different tagset than emMorph or Universal Dependencies (UD).

In our research, we used two neural morphological generators [6] (an emMorph and an UD model) that was trained by the Hungarian Research Centre for Linguistics.

3. Corpora

Names corpus: The names were gathered from the list of names that can be registered as given names [12]. This list is updated monthly and the local version of the corpus was cached in October of 2022. The corpus itself is divided into two parts, male and female given names. This distinction is important, because certain gender implying words, such as father, mother etc. in the sentence can be connected to the swapped name and in many cases the gender of the word can disambiguate the context for the conversation.

Family names: Unfortunately no ready made corpora are available of Hungarian family names, Only the most used one hundred [5] second names are published. As a fall back if an unknown family name is present the replacing algorithm was tasked to replace the unknown name with a male name rather than guessing a family name.

(1) a. ***Maga***
FORMAL PERSONAL PRONOUN THIRD PERSON SINGULAR

‘Oneself’

b. ***Fodor Kriszta***
NAME THAT CAN BE INTERPRETED AS TWO GIVEN NAMES

c. ***Budai Krisz***
NICKNAME

d. ***Pálma Veresné Berta***
OUT OF ORDER NAME

e. ***Bihar Megyei Tanács***
NAME OF AN INSTITUTION

f. ***BRÜLL ADÉLNAK***
ALL CAPITAL NAME

g. ***Aranycsapat***
NAME OF A COLLECTIVE

Evaluation corpora: Three texts were chosen for evaluating the pipeline. All of them are contemporary variations of the language and are in line with the use cases connected to journalism and the media:

- reporting on social media: “comments”: scraped comments from a social media post, especially hard text for its noisy nature
- reporting on legal documents “Ady letters”: letters from Endre Ady, selected for the usage of capitalized pronouns (1a) and full capital names (1f)
- official reporting “spok”: excerpt from MNSZ [21]
- reporting on historical documents “huwiki”: mixed sentences from the Hungarian Wikipedia
- reporting on not spoken news “newscrawl”: crawled news from reputable journalists collected in NerKor [17]

These corpora contained many opportunities for the models to fail: capitalized personal pronouns (1a), reverse ordered names (1d), nicknames (1c), names that have given names as first name (1b), all capitalized names (1f) names of collective (1g), place and institution names (1e).

4. Pipeline architecture and modules

4.1. Modules

In our pipeline we have five main modules:

1. **Preprocessing:** For preprocessing, HuSpaCy and emMorph was used. The input text is splitted into sentences and tokenized.
2. **Morphological analyzer**: HuSpaCy and emMorph modules were integrated. Morphological analysis can be performed with either HuSpaCy in UD format or EmMorph in emMorph code format.
3. **NER:** For named entity recognition, a fine-tuned huBERT model [22] was used. The model was tested on NerKor that achieved 90.04 precision, 91.17 recall and 90.60 F1 scores.
4. **Name databases:** The names were collected from the official list of Hungarian surnames that are recognised [12] and the first one hundred most used family names [5].
5. **Morphology genarator:** Two neural-based Hungarian morphology generators were used that provided by the Hungarian Research Centre for Linguistics [6]. One emMorph and one UD morphology generator. The generator models were trained with Marian neural machine translation system. The eMagyar model could gain more than 96% accuracy and the UD model could achieve more than 94% accuracy.

4.2. Architecture

The pipeline serves as an integration of various modules, which are introduced in the Modules section. This section provides an overview of the pipeline architecture. As illustrated in Figure 1, a module may be utilized multiple times throughout the process. For example, the morphological analyzers are employed for both tokenization of the raw text and subsequent morphological analysis.

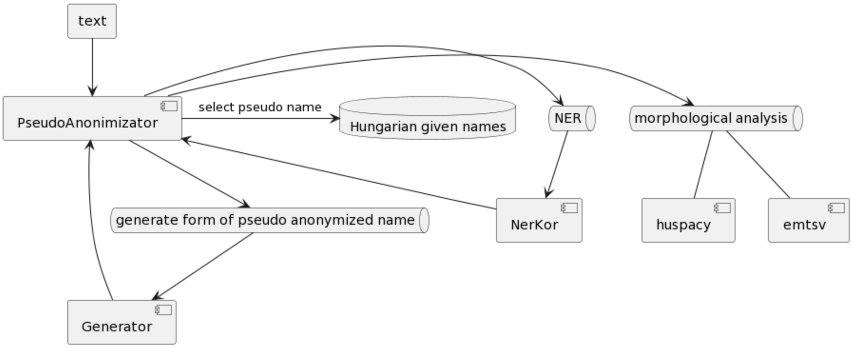

Figure 1. Structure diagram of the pseudononymization Tool.

As illustrated in Figure 2, the anonymization process is inherently sequential and challenging to parallelize. The names must first be identified, followed by selecting suitable replacement names, and finally performing the name swapping. This swapping process introduces a conundrum: not only is the identification of names potentially inaccurate, but the original and new names may also have different lengths, requiring the delta of lengths to be stored alongside with the mapping of the name to the pseudonym which will be mapped in the text.

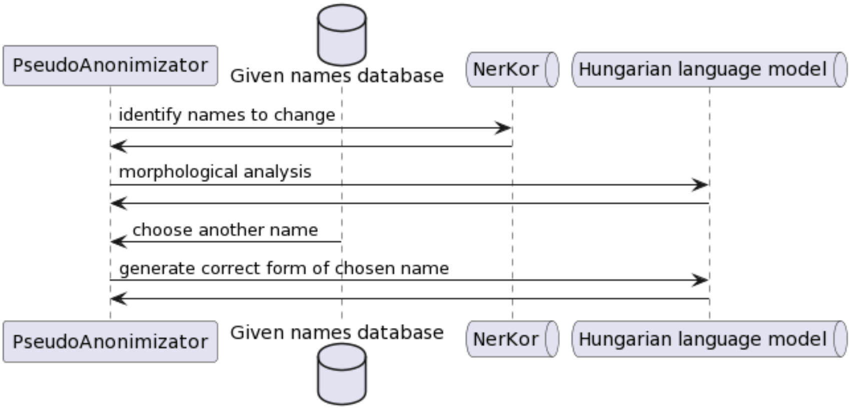

Figure 2. Activity diagram of the pseudononymization Tool.

Furthermore, neural models typically perform best when processing smaller text portions, similar to the complete sentences they were trained on. Consequently, the input text is first divided into sentences, which are then processed in a paginated manner using the NER system to identify potential named entities and their positions.

5. Results and evaluation

In our first experiment, we evaluated the performance of our NER model on person names. Using NerKor corpus, the NER model achieved 96.25 F1 score on person names (words tagged with [PER]).

In our second experiment, we evaluated the performance of eMagyar and UD morphology generators on named entities using the NerKor corpus. We collected the morphologically analyzed tokens from NerKor, filtered and extracted the unique names, and applied morphology generators to these names. The results of our evaluation are presented in Table 1. The 'all' column shows the performance of the models on all names, but this result may be biased as nearly 86% (eMagyar: 85.62%; UD: 86.07%) of names are in nominative case. Therefore, we also evaluated the performance of the models on non-nominative cases, which is shown in the 'filtered' column.

Table 1. Performance of morphology generators on named entities.

	all	filtered
emMorph	95.10%	84.51%
UD	92.85%	79.39%

For further evaluation we have established an ideal scenario, when all names are successfully replaced and no false positives are present, and a worst case, when no names are recognized and every other word gets replaced but the names. These scenarios are significantly distinct from either the eMagyar or the huSpacy morphological analyzers when compared with Student's t-test [18]. The evaluation metrics used are as follows:

1. True Positive: The number of actual named entities (real names) found by the pipeline.

2. False Positive: The number of incorrect named entities identified by the pipeline that are not real names.

3. False Negative: The number of real named entities that were not identified by the pipeline.

4. All: The total number of named entities (both correctly identified and incorrectly identified) in the text.

The p-values indicate the probability of obtaining the observed results if there is no significant difference between the pipelines. The hypothesis being tested is whether there is a significant difference in performance between the two pipelines. In the spok corpus with the eMagyar analyzer the distinction from the worst ($p = 0.0001$) and the ideal ($p = 0.0029$) are hairly distinct compared to the huSpacy analyzed which is not significantly distinct ($p = 0.5933$). The same can be observed with the analysis of the two other corpora. However, the similarity of the same pipeline on different texts is not as close as one might think eMagyar "spok" and "comments" are not close ($p = 0.0001$), only the Ady letter text and the spok corpus with the huSpacy pipeline was close ($p = 0.0454$), but they still differed significantly.

The pipelines yielded results with many false positives, such as reporting a name in the middle of a word or finding the name in two separate parts, but in such a way that the two findings are next to each other without even a character of difference. In such cases, as a remedy, a unification algorithm was used, which extended the first hit and deleted the second. Even with this measure, the false positive percentage of all found names is high, especially in the "comments" corpus (Table 2).

Table 2. Performance of pseudo anonymization or eMagyar morphological analyzer.

text	True Positive	False Positive	False Negative	All
spok	7	8	2	13
comments	42	25	9	67
Ady letters	11	4	9	20
huwiki	85	2	0	87
newscrawl	103	2	0	105

Table 3. Performance of pseudo anonymization or huSpacy morphological analyzer.

text	True Positive	False Positive	False Negative	All
spok	6	5	5	11
comments	58	19	3	77
Ady letters	6	5	9	15
huwiki	85	0	4	86
newscrawl	103	2	1	128

When analyzing the data from the huwiki and newscrawl corpora, both are collected from NerKor [17], it is evident that the false positive rates for these datasets are generally lower than those observed in the "comments" corpus. For instance, the eMagyar pipeline had only 2 false positives in both huwiki and newscrawl corpora, while the huSpacy pipeline showed no false positives in the huwiki corpus and only 2 in the newscrawl corpus (Tables 2 and 3). This suggests that the performance of both pipelines in terms of false positives may vary depending on the specific corpus being analyzed, and further research or fine-tuning may be required to optimize the pipelines for each individual corpus.

Table 4. Performance of NerKor with different tokenization pipelines.

corpus	pipeline	All Positives (True + False)	All real names	All words
spok	emagyar	24	10	735
spok	HuSpacy	24	10	735
comments	emagyar	128	67	876
comments	huspacy	126	67	876
Ady letters	emagyar	37	20	1369
Ady letters	huspacy	32	20	1369
huwiki	emagyar	85	87	2785
huwiki	huspacy	85	86	2785
newscrawl	emagyar	103	105	4551
newscrawl	huspacy	103	128	4551

Table 4 shows the performance of NerKor with two tokenization pipelines on different datasets. In general, both pipelines perform similarly, identifying named entities in the corpora. However, in the “newscrawl” dataset, the “Huspacy” pipeline outperforms “emagyar” in identifying named entities (128 vs. 103). The total word count remains consistent across all corpora and tokenization pipelines.

(2) a. *751 SCHÖPFLIN* ***ALADÁRNAK***
751 SCHÖPFLIN **to Aladar**
751 Schöpflin **Ilájnak**
751 Schöpflin **to Ilja**

b. *Édes, Drága, áldott* ***Adélom,***
Sweet, Precious, blessed **my Adel**,
Édes, Drága, áldott **Darlám**,
Sweet, Precious, blessed **my Darla**,

c. *ima által kapcsolatba lépünk* ***Istennel,***
prayer through get connected **with God**,
ima által kapcsolatba lépünk **Gyárfással**,
prayer through get connected **with Gyárfás**,

In Example 2 the proper form generation can be observed in action. These examples are from the ‘Ady letters’ corpus and they were generated from form tokenized and analyzed by the eMagyar pipeline. As you can see the case stayed the same and even the sound assimilation was generated properly. The last example showcases a certain error that poses a great challenge whereas the NER system identifies collectives or supernatural beings as persons and treats them with great respect regarding their privacy.

> ‘**Benjámin** Somogyi’
>
> ‘**Benjámin** Somogyi’
>
> ‘Magatokon lehetne a legtöbbet spórolni, de arról hallani sem akartok! Az egyes élelmiszer hatósági árát, a kereskedőknek kell köszönni, nem nektek!,’
>
> ‘It would be possible to save the most on yourselves, but you don’t even want to hear about it! The regulatory price of individual food products should be credited to the traders, not you!’
>
> ‘Hegedűsné Krizsák **Barbara**’,
>
> ‘Mrs. Hegedűsné Krizsák **Barbara**’,
>
> ‘**Benjámin** Somogyi így igaz.’,
>
> ‘**Benjámin** Somogyi that’s true.’,
>
> ‘Száváné **nagy Balzsam**’,
>
> ‘Száváné **Nagy Balzsam**’,
>
> ‘**Benjámin** Somogyi A kereskedők nem maguktól találták ezt ki. Majd hülyék lennének maguk ellen dolgozni.’
>
> ‘**Benjámin** Somogyi The traders didn’t come up with this on their own. They would be fools to work against themselves.’

The previous quote demonstrates the pipeline in action on the comments corpus. One of the main features is the consistency of names. All the given names were typeset bold for the quote to show how the same person, who should be known as Benjámin according to the pipeline and wears the family name of Somogyi, is consistently swapped to be Benjámin. However, an error is also present in this presentation. The family name “Nagy” is mistakenly replaced with all lower case and part of the tagging is also present for now, as this error is under investigation.

6. Conclusion

We have presented a pseudononymization pipeline with a command line interface and a web service, tailored for the Hungarian language. All the tools used in the pipeline, as well as the pipeline itself, are freely available and open source. Instructions for local deployment can be found in the git repository, ensuring easy reproducibility. Our approach achieved impressive results as the first of its kind, extending beyond traditional use cases in medicine and legal documents. We believe that anonymization is equally important in media and business contexts, where noise is more prevalent than in the aforementioned fields. The testing corpora posed significant challenges, yet our models were able to effectively handle them, demonstrating the potential for broader applications of our pseudononymization approach.

7. Further work

In the future we want to expand this tool with the ability to create a database with which the pseudonymized text can be deanonymized with ease, this should be provided with a switch to the command line interface and a separate endpoint for the web server. Additionally more models could be incorporated to expand the possibility to find the best model for this task.

References

[1] H. Aamot, C. D. Kohl, D. Richter, P. Knaup-Gregori: *Pseudonymization of patient identifiers for translational research*, BMC Medical Informatics and Decision Making 13 (2013), pp. 1472–6947.

[2] H. Dalianis: *Pseudonymisation of Swedish Electronic Patient Records Using a Rule-Based Approach*, in: Proceedings of the Workshop on NLP and Pseudonymisation, Turku, Finland: Linköping Electronic Press, Sept. 2019, pp. 16–23, url: https://aclanthology.org/W19-6503.

[3] European Commission: *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance)*, 2016, url: https://eur-lex.europa.eu/eli/reg/2016/679/oj.

[4] M. Hulden: *Foma: a Finite-State Compiler and Library*, in: Proceedings of the Demonstrations Session at EACL 2009, Athens, Greece: Association for Computational Linguistics, Apr. 2009, pp. 29–32, url: https://aclanthology.org/E09-2008.

[5] M. of Interior Deputy State Secretariat for Data Registers: *Most common 100 family names*, data retrieved from Ministry of Interior Deputy State Secretariat for Data Registers, https://www.nyilvantarto.hu/letoltes/statisztikak/kozerdeku_csaladnev_2022.xlsx, 2022.

[6] L. J. Laki, N. Ligeti-Nagy, N. Vadász, Z. Gy. Yang: *Neural Morphological Generators for Hungarian*, in: XIX. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2023), Szeged, Hungary: Szegedi Tudományegyetem, Informatikai Intézet, 2023, pp. 331–340.

[7] K. Mátyus: *Anonimizálási gyakorlat?*, in: IX. Magyar Számítógépes Nyelvészeti Konferencia, Szeged: Szegedi Tudományegyetem, 2013, pp. 338–342.

[8] D. M. Nemeskey: *Egy emBERT próbáló feladat*, in: XVI. Magyar Számítógépes Nyelvészeti Konferencia, Szeged: Szegedi Tudományegyetem, 2020, pp. 409–418.

[9] D. M. Nemeskey: *Introducing huBERT*, in: XVII. Magyar Számítógépes Nyelvészeti Konferencia, Szeged, Magyarország: Szegedi Tudományegyetem, Informatikai Intézet, 2021, pp. 3–14.

[10] A. Novák: *A New Form of Humor – Mapping Constraint-Based Computational Morphologies to a Finite-State Representation*, in: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), ed. by N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, S. Piperidis, Reykjavik, Iceland: European Language Resources Association (ELRA), May 2014, isbn: 978-2-9517408-8-4.

[11] A. Novák, B. Siklósi, Ch. Oravecz: *A New Integrated Open-source Morphological Analyzer for Hungarian*, in: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), ed. by N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis, Portorož, Slovenia: European Language Resources Association (ELRA), May 2016, isbn: 978-2-9517408-9-1.

[12] N. K. Nyelvművelő és Nyelvi Tanácsadó Kutatócsoport: *Bejegyzésre alkalmasnak minősített utónevek jegyzéke*, 2022, url: http://www.nytud.hu/oszt/nyelvmuvelo/utonevek/index.html (visited on 10/25/2022).

[13] Gy. Orosz, A. Novák: *PurePos 2.0: a hybrid tool for morphological disambiguation*, in: Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013, Hissar, Bulgaria: INCOMA Ltd. Shoumen, BULGARIA, Sept. 2013, pp. 539–545, url: https://aclanthology.org/R13-1071.

[14] Gy. Orosz, Z. Szántó, P. Berkecz, G. Szabó, R. Farkas: *HuSpaCy: an industrial-strength Hungarian natural language processing toolkit*, Szeged, 2022.

[15] B. Riedl, V. Grascher, S. Fenz, T. Neubauer: *Pseudonymization for improving the Privacy in E-Health Applications*, in: Proceedings of the 41st Annual Hawaii International Conference on System Sciences (HICSS 2008), 2008, pp. 255–255, doi: 10.1109/HICSS.2008.366.

[16] B. Schütze: *Use of medical treatment data outside of the patient supply: best way pseudonymisation*, Dtsch Med Wochenschr 137(16) (2012), pp. 844–850.

[17] E. Simon, N. Vadász: *Introducing NYTK-NerKor, A Gold Standard Hungarian Named Entity Annotated Corpus*, in: Text, Speech, and Dialogue - 24th International Conference, TSD 2021, Olomouc, Czech Republic, September 6-9, 2021, Proceedings, ed. by K. Ekstein, F. Pártl, M. Konopík, vol. 12848, Lecture Notes in Computer Science, Springer, 2021, pp. 222–234, doi: 10.1007/978-3-030-83527-9_19.

[18] Student: *The probable error of a mean*, Biometrika (1908), pp. 1–25.

[19] V. Trón, G. Gyepesi, P. Halácsy, A. Kornai, L. Németh, D. Varga: *Hunmorph: open source word analysis*, in: Proceedings of the ACL 2005 Software Workshop, ed. by M. Jansche, Ann Arbor: ACL, 2005, pp. 77–85.

[20] V. Trón, P. Halácsy, P. Rebrus, A. Rung, P. Vajda, E. Simon: *Morphdb.hu: Hungarian lexical database and morphological grammar*, in: Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06), Genoa, Italy: European Language Resources Association (ELRA), May 2006, url: http://www.lrec-conf.org/proceedings/lrec2006/pdf/683_pdf.pdf.

[21] T. Váradi: *The Hungarian National Corpus*, in: Proceedings of the Third International Conference on Language Resources and Evaluation (LREC-2002), Las Palmas de Gran Canaria: European Language Resources Association, 2002, pp. 385–389.

[22] Z. Gy. Yang, T. Váradi: *Training language models with low resources: RoBERTa, BART and ELECTRA experimental models for Hungarian*, in: Proceedings of 12th IEEE International Conference on Cognitive Infocommunications (CogInfoCom 2021), Online: IEEE, 2021, pp. 279–285.