sciendo
DOI 10.2478/jazcas-2025-0017

MasKIT – ANONYMIZATION AND PSEUDONYMIZATION OF CZECH LEGAL TEXTS

JIŘÍ MÍROVSKÝ[1] – TEREZA NOVOTNÁ[2] – BARBORA HLADKÁ[3]
[1]Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic (ORCID: 0000-0003-2741-1347)
[2]Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic (ORCID: 0000-0002-1426-4547)
[3]Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic (ORCID: 0000-0003-4950-4587)

MÍROVSKÝ, Jiří – NOVOTNÁ, Tereza – HLADKÁ, Barbora: MasKIT – Anonymization and Pseudonymization of Czech Legal Texts. Journal of Linguistics, 2025, Vol. 76, No 1, pp. 192 – 199.

Abstract: MasKIT is a command-line tool, an on-line web application and a REST API service for anonymization and pseudonymization of Czech legal texts. Taking a plain text as input (e.g. a letter sent by a legal authority to a citizen), it runs external services for dependency parsing and named entity recognition and then via a rule-based approach identifies and replaces sensitive information in the text.

Keywords: anonymization, pseudonymization, legal texts, tool, web interface, REST API, Czech

1 INTRODUCTION

1.1 Motivation

Courts in most countries are now obliged to publish their decisions online. Access to case law is a cornerstone of access to justice, which is one of the fundamental principles of the rule of law. Thus, court decisions serve not only as precedents for subsequent cases, but also as a source of a great deal of legal information, interpretations of legal norms or even information about the functioning of society and various political changes. Such analyses then require access to case law as a whole, i.e. full texts and underlying metadata.

Opposite to access to justice in this case is the protection of privacy, which is also a constitutionally guaranteed right. The invasion of privacy caused by the publication of personal data in decisions can be severe, especially in the case of criminal decisions, and can even lead to secondary victimisation and endangerment of victims of crime. The precise anonymization of published judicial decisions is a key element in protecting the privacy of litigants. At the same time, anonymization is also a tool to enable the publication of case law to the general public and thereby strengthen access to justice.

Anonymization is therefore an important step, a *conditio sine qua non*, in the process of publishing legal documents (e.g. court or administrative decisions), but at the same time it must be carried out very precisely and in accordance with the legislation governing the protection of the privacy and personal data of the subjects. Automated anonymization tools therefore have the great potential to save significant time and manual labour in the courts. However, such a tool must be sufficiently precise, comply with data protection legislation and ideally meet the transparency and explainability requirements of both national and European regulations.

In the Czech legal environment, the scope of anonymized data and the method of anonymization are governed by Decree No. 403/2022 Coll., on the publication of court decisions, which implements Act No. 6/2002 Coll., on courts and judges.

In this paper, we present MasKIT, an open-source tool for automatic anonymization and pseudonymization of Czech legal documents. The tool is being developed to comply with the legislative anonymization rules mentioned above. At the same time, it is available under open licences and widely usable not only for anonymization/pseudonymization of legal documents in courts and public administration, but also for the public. The methods on which MasKIT is based meet the strict conditions for transparency and explainability of processes imposed by European legislation (AI Act, GDPR Art. 22).

In the rest of the Introduction, we present recent related work on anonymization/ pseudonymization. In Section 2, we describe the system architecture and give a step-by-step example of processing a Czech sentence both in the anonymization and pseudonymization modes. In Section 3, the user interface is shortly introduced. Section 4 is dedicated to evaluating the system and concluding the paper.

1.2 Related work

Csányi et al. (2021) discusses the complexities of anonymizing legal documents, highlighting the need to balance privacy protection with the preservation of information utility. It emphasizes that while Named Entity Recognition methods are crucial, they are insufficient on their own. The authors advocate for integrating machine learning techniques with anonymization models, such as differential privacy, to effectively reduce re-identification risks.

Oksanen et al. (2022) introduces the ANOPPI tool developed for (semi-) automatic anonymization of Finnish texts. The tool can be used both as a web application and programmatically through a REST API. Evaluation shows that ANOPPI performs well with different types of documents, however, further improving the performance of the named entity recognition and disambiguation methods would enhance the usefulness of the software. The tool is being published as open source for public use by the Ministry of Justice in Finland.

Glaser et al. (2021) used the BERT architecture to train an anonymization model that takes into account the context of the anonymized data. Such a method then, according to the authors, does not require non-anonymized data to train, but can be applied to anonymized publicly available legal documents.

Similarly, Licari et al. (2022) presented a model to anonymize Italian judicial decisions, based on transformers and spacy entity recognition.

Recently, the anonymization of legal texts has begun to combine traditional rule-based approaches with fine-tuning and domain-specific pre-training of large language models. For example, Niklaus at al. (2023) explores improvements to the anonymization system used by the Swiss Federal Supreme Court, known as Anom2. The study focuses on enhancing the identification and masking of personal information in legal texts by integrating machine learning techniques. The researchers compiled a large annotated dataset containing entities requiring anonymization, which served as a training and evaluation resource. By pre-training BERT-based models on domain-specific legal data, they achieved an F1-score improvement of over 5% compared to models trained without such in-domain knowledge.

2 THE SYSTEM DESCRIPTION

2.1 The system architecture

MasKIT is a command-line tool and also a client-server application with a web client interface and a REST API server. The tool (the server) is written in Perl and calls two state-of-the-art external services to pre-process the texts:

- UDPipe (Straka 2018) for syntactic analysis of the text, and
- NameTag (Straková et al. 2019) for recognition of named entities.

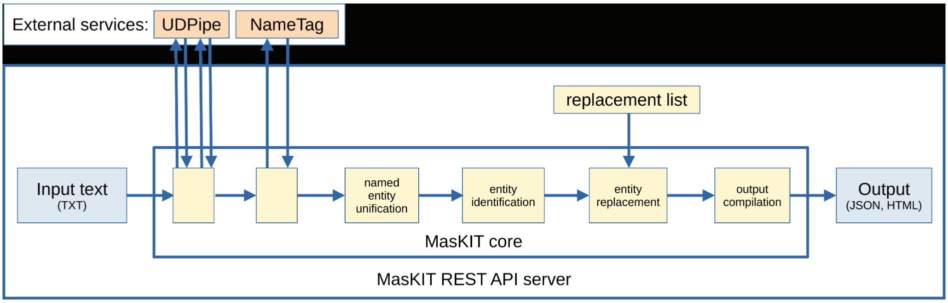


Fig. 1. MasKIT application architecture

The subsequent rule-based analysis uses information from both the external tools. The dataflow of the whole process is given in Fig. 1. The system accepts a plain text as the input and performs the following steps:

(1) **Morphological and syntactic analysis:** UDPipe is called twice with two different models: First, an older model that gives preference to longer sentences is

called for sentence segmentation, which is suitable for legal texts that often contain addresses and various abbreviations with dots in the middle of sentences (wrong segmentation at these places would later harm the performance of named entities recognition). Second, the pre-segmented text is passed to a new model to obtain state-of-the-art morphological tagging and syntactic parsing in the Universal Dependencies framework.[1]

(2) **Named entities recognition:** The parsed data are sent to NameTag for recognition of named entities. Because of time limitations of the NameTag service, long texts are split to shorter segments and processed separately.

(3) **Named entities unification:** As NameTag sometimes fails to recognize all occurrences of the same named entity in a single document, or sometimes mis-classifies some of the occurrences,[2] unification of classification of single-word named entities is performed. For each token[3] in the text that is not a part of a multi-word named entity, all named-entity marks are counted. Unless most of the occurrences of the token are unmarked (without an assigned named-entity mark), all the tokens are (re-)assigned the most frequent named-entity mark. If there are more than one most frequent mark, the class relevant for MasKIT is preferred.[4]

(4) **Rule-based analysis:** The text is analyzed sentence by sentence and token by token, traversing the dependency syntax trees. Expressions that should be anonymized are detected with a series of manually encoded rules, taking advantage of the assigned named-entity marks and the parsed dependency tree structure.

Some types of expressions are best detected using dedicated Perl libraries, which give more reliable results than the named-entity marks (e.g. e-mail addresses). Some hard-to-parse expressions are best detected directly from the surface form of the sentence (such as agenda reference numbers), but for many of the detected types of expressions, using the parsed dependency tree structure is beneficial: For example, detecting dates of birth relies on finding key lemmas (such as *narození* ‘birth’ or *narodit (se)* ‘be born’ in the correct dependency relation with a date, regardless of other words appearing in the surface word order.

(5) **Output generation:** After the whole text is processed, the output is produced in the selected output format (TXT, HTML, CoNLL-U).[5]

[1] https://universaldependencies.org/

[2] Further study would be required to show if and how much it is related to processing longer texts in segments.

[3] More precisely: a combination of its lemma and part of speech.

[4] E.g. a ‘country name’ mark is not relevant for MasKIT, as countries do not get anonymized.

[5] The CoNLL-U format is only available from the command line (not in the web interface or via REST API).

2.2 Anonymization vs. pseudonymization

MasKIT supports both anonymization and pseudonymization. In the anonymization mode, the sensitive expressions are replaced by their classes, while in the pseudonymization mode random words of the same class are used. Let us assume that the sentence from Example (1) is entered in the system.

(1) Paní Marie Nováková z Myslíkovy ulice č. 25 dostala dopis od firmy Škoda Auto, a.s..
[Mrs. Marie Nováková from Myslíkova street No. 25 received a letter from the company Škoda, a.s..]

In the anonymization mode, a woman's surname *Nováková* is replaced by class *M-ŽENA-PŘÍJMENÍ-1* 'M-WOMAN-SURNAME-1', where the numeric index (here *1*) distinguishes replacements for different surnames, see Example (2).

(2) Paní **M-ŽENA-JMÉNO-1** **M-ŽENA-PŘÍJMENÍ-1** z **M-ULICE-1** ulice č. **M-ČÍSLO-ULICE-1** dostala dopis od firmy **M-FIRMA-1**.
[Mrs. **M-WOMAN-NAME-1** **M-WOMAN-SURNAME-1** from **M-STREET-1** street No. **M-STREET-NUMBER-1** received a letter from the company **M-COMPANY-1**.]

In the pseudonymization mode, one of twenty pre-defined surname replacements is used (e.g. *Pospíšilová*) in the correct morphological case, see Example (3).

(3) Paní **Alena** **Pospíšilová** z **Květinové** ulice č. **43** dostala dopis od firmy **UniTechna**.
[Mrs. **Alena** **Pospíšilová** from **Květinová** street No. **43** received a letter from the company **UniTechna**.]

Further occurrences of surname *Nováková* in the same text get replaced by the class with the same index, or with the same surname replacement. Male and female surnames are tied, so if, e.g. also *Mr. Novák* appeared in the text, it would be replaced with *Mr. M-MUŽ-PŘÍJMENÍ-1* or *Mr. Pospíšil*, respectively, to match the corresponding female surname.

The system always tries to replace a multiple-word expression (such as *Škoda Auto, a.s.*) as a whole, i.e. the whole company name is replaced by a single *M-FIRMA-1* 'M-COMPANY-1' or *UniTechna*, resp.

For inspection of the result in comparison with the original text, the user can optionally display also the original expressions in subscript next to the anonymized/ pseudonymized replacements, see Example (4) and also Fig. 2.

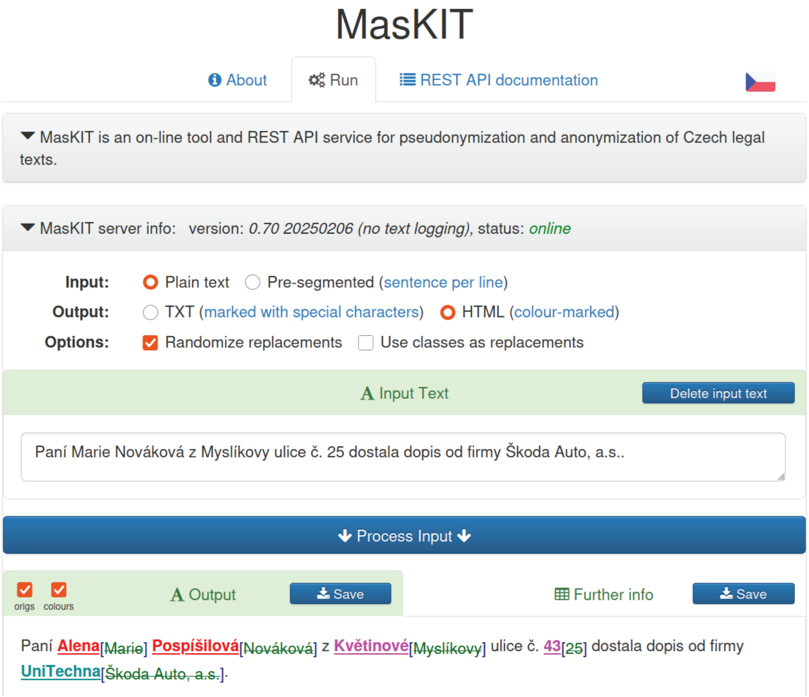
(4) Paní **Alena**[~~Marie~~] **Pospíšilová**[~~Nováková~~] z **Květinové**[~~Myslíkovy~~] ulice č. **43**[~~25~~] dostala dopis od firmy **UniTechna**[~~Škoda Auto, a.s.~~].
[Mrs. **Alena**[~~Marie~~] **Pospíšilová**[~~Nováková~~] from **Květinová**[~~Myslíkova~~] street No. **43**[~~25~~] received a letter from the company **UniTechna**[~~Škoda Auto, a.s.~~].]

3 USER INTERFACE

The MasKIT server can be either run as a command-line utility,[6] or it can be accessed via a web client or a REST API service.

3.1 Web interface

The MasKIT web client is written in PHP[7] and Bootstrap 3[8] and provides a browser-based interface to the server. The user enters a text (directly or as a file) and submits it. The text is passed to the server via REST API, processed by the server and the result is then presented to the user in an interactive way, see Fig. 2.



MasKIT is an on-line tool and REST API service for pseudonymization and anonymization of Czech legal texts.

MasKIT server info: version: *0.70 20250206 (no text logging)*, status: *online*

Input: ● Plain text ○ Pre-segmented (sentence per line)

Output: ○ TXT (marked with special characters) ● HTML (colour-marked)

Options: ☑ Randomize replacements □ Use classes as replacements

Paní Marie Nováková z Myslíkovy ulice č. 25 dostala dopis od firmy Škoda Auto, a.s..

Paní Alena[~~Marie~~] Pospíšilová[~~Nováková~~] z Květinové[~~Myslíkovy~~] ulice č. 43[25] dostala dopis od firmy UniTechna[~~Škoda Auto, a.s.~~].

Fig. 2. MasKIT web interface

[6] See the program documentation: https://ufal.mff.cuni.cz/maskit/users-manual.
[7] https://www.php.net/
[8] https://getbootstrap.com/

3.2 Public REST API anonymous server

The Institute of Formal and Applied Linguistics at Charles University is running a publicly available MasKIT web client[9] and REST API service,[10] free for non-commercial usage. As the submitted texts may be of sensitive nature, the server does not store any part of the processed text. The same anonymity is guaranteed for requests from this server by the external services UDPipe and NameTag.[11]

4 EVALUATION AND CONCLUSION

To evaluate MasKIT, 7 annotators (students of the Faculty of Law at Masaryk University, Brno) have manually annotated 53 legal documents (5,373 sentences, 127 thousand tokens) of the following nature: court decisions (9), citizen's letters to an authority (2), authority's decisions (21), legal advices (12), lawsuits/appeals (3), ombudsman's reports (6). The annotators marked 2,372 sensitive expressions to be anonymized, classifying each occurrence in one of 20 classes recognized by MasKIT plus category 'other' for any other type (not supported by MasKIT). In the same documents, MasKIT anonymized and classified the sensitive information with Recall of 0.8, Precision of 0.64 and F1 measure of 0.7 on recognition of expressions to be anonymized. The classification accuracy on expressions marked both by the annotators and MasKIT was 0.91. The Precision and Recall are comparable to results in Glaser et al. (2021): they evaluated their system on 46 documents from Munich district and financial courts and reported Precision in the range from 0.63 to 0.69 and Recall in the range from 0.52 to 0.79. Their classification accuracy was 0.73.

MasKIT is still under development and does not yet support all classes of sensitive information as defined in Decree No. 403/2022 Coll., on the publication of court decisions. As all these classes have been included in our evaluation, improvements in comparison with the currently reported results are to be expected in future versions. In the present version, MasKIT complies with approx. 20 out of 35 categories designed to be anonymized by the Degree, not yet implementing, most of all, academic titles, account numbers, payment variable symbols, data box numbers and personal ID numbers.

ACKNOWLEDGEMENTS

The work was supported by the project "PONK – Computer-Assisted Clear Legal Writing" (reg. no. TQ01000526, co-funded by the Technology Agency of the Czech Republic) and the project "Human-centred AI for a Sustainable and Adaptive

[9] https://quest.ms.mff.cuni.cz/maskit/

[10] See MasKIT REST API documentation: https://ufal.mff.cuni.cz/maskit/api-reference.

[11] The only information that may be logged: time of usage, size of the processed data, the system configuration and the IP address from where the MasKIT service is accessed.

Society” (reg. no. CZ.02.01.01/00/23_025/0008691, co-funded by the European Union). The research reported in the present contribution has been using language resources developed, stored and distributed by the LINDAT/CLARIAH-CZ project of the Ministry of Education, Youth and Sports of the Czech Republic (LM2023062).

References

Csányi, G. M., Nagy, D., Vági, R., Vadász, J. P., and Orosz, T. (2021). Challenges and Open Problems of Legal Document Anonymization. Symmetry, 13(8), 1490. Accessible at: https://doi.org/10.3390/sym13081490.

Glaser, I., Schamberger, T., and Matthes, F. (2021). Anonymization of German legal court rulings. New York, NY, USA: Association for Computing Machinery. ICAIL 21. Accessible at: https://doi.org/10.1145/3462757.3466087.

Niklaus, J., Mamié, R., Stürmer, M., Brunner, D., and Gygli, M. (2023). Automatic Anonymization of Swiss Federal Supreme Court Rulings. In Proceedings of the Natural Legal Language Processing Workshop 2023, pp. 159–165, Singapore. Association for Computational Linguistics. Accessible at: https://doi.org/10.18653/v1/2023.nllp-1.16.

Licari, D., Romano, M., and Comande, G. (2022). Automatic Anonymization of Italian Legal Textual Documents using Deep Learning. ITA 2022. Accessible at: https://www.iris.ss-sup.it/handle/11382/548773.

Oksanen, A., Hyvönen, E., Tamper, M., Tuominen, J., Ylimaa, H., Löytynoja, K., Kokkonen, M., and Hietanen, A. (2022). An Anonymization Tool for Open Data Publication of Legal Documents. In Joint Proceedings of ISWC2022 Workshops: the International Workshop on Artificial Intelligence Technologies for Legal Documents (AI4LEGAL) and the International Workshop on Knowledge Graph Summarization (KGSum), pp. 12–21.

Straka, M. (2018). UDPipe 2.0 Prototype at CoNLL 2018 UD Shared Task. In Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. Association for Computational Linguistics, Brussels, Belgium, pp. 197–207. Accessible at: https://doi.org/10.18653/v1/K18-2020.

Straková, J., Straka, M., and Hajič, J. (2019). Neural Architectures for Nested NER through Linearization. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Florence, Italy, pp. 5326–5331. Accessible at: https://doi.org/10.18653/v1/P19-1527.