Rules-based system for Czech legal text readability

Kateřina Motalík Hodková and Ivan Kraus and Barbora Hladká
Institute of Formal and Applied Linguistics
Faculty of Mathematics and Physics, Charles University
Prague, Czech Republic
{hodkova, hladka}@ufal.mff.cuni.cz, ivakra@centrum.cz

Abstract

In this paper, we present a set of linguistic rules, employed to enhance the readability of legal texts. The rules were compiled and implemented as a rule-based module of PONK, an advisory tool that contributes to simplification and higher clarity of Czech legal texts, especially those intended for non-expert audience. Based on recurring phenomena in authentic texts and relevant scientific sources, the rules cover mainly the domains of syntax and lexicon. In addition, we present the results of application of the rules to a corpus of authentic legal texts, evaluated by a human annotator, and examine their impact.

1 Introduction

Legal texts can pose a significant challenge, particularly for readers without legal training. Their complexity arises from several factors, including long and complicated syntactic structures, legal terminology, and the use of archaic, ambiguous, or semantically vague expressions. Although such features can often be linked to historical or stylistic conventions of legal drafting, they can substantially hinder comprehension, and individuals affected by legal documents (e.g., citizens bound by laws of a country, parties of a contract, or parties of a court process) may struggle to interpret legal texts accurately, which may result in grave consequences. The readability and clarity of legal texts, therefore, presents an important objective, both from the perspective of legal certainty and from the general perspective of accessibility and transparency.

As the linguistic obstacles of legal texts are strongly language-dependent and potentially related to specific drafting conventions, we decided to focus on legal texts written in Czech. The paper presents both the design of the rule set (in the context of Czech grammar) and its application to authentic Czech legal texts, which is incorporated within the PONK tool. PONK tool is an advisory tool designed to identify difficult-to-understand passages of legal texts (the tool does not automatically generate simplified versions of the texts). To assess the impact of the proposed rules, the identified segments are subjected to human annotation. Its purpose is to judge whether modifications of text segments indicated by PONK tool would contribute to an increased readability of the text or not. This annotation allows us to estimate the practical effectiveness of the rules and identify areas for further refinement.

In addition, our aim is to formulate the rules and the overall approach in a way that allows the methodology to be adapted to other languages with appropriate linguistic adjustments.

2 Legal text and related works

Legal language is a type of language for special purposes belonging to the standard register of a given language (Hodková et al., 2021). It is the language of legal text (written or spoken) and its objective is to transfer information belonging to the legal domain. Among the specialty languages, legal language occupies a particular position: whereas most specialty languages (be it in the domain of biology, astrophysics, or medicine) concern mostly only the given group of experts, legal language and legal texts involve virtually everyone via laws, contracts, etc. According to Baldinger (1984), specialty languages are, unlike natural languages, motivated and not arbitrary.

Terminological units constitute an inherent part of legal texts. Terms are lexical units designating legal concepts, abstract semantic units belonging to the domain of law (Cornu, 2005; Tomášek, 2003). Legal concepts often represent an obstacle in understanding legal text (mainly for readers without legal education). However, given that legal concepts are precisely defined (whether it is a legal definition or legal citation, Hodková et al. (2021)), they cannot

Proceedings of the 20th Linguistic Annotation Workshop (LAW XX), pages 22–32

be altered, because this could obscure the intended meaning.

In anglophone countries, the tendencies to simplify the language of legal texts have a long tradition, concerning texts from jury instructions to contracts (Charrow and Charrow, 1979; Chromá, 2016; Diamond et al., 2012; Hartig and Lu, 2014; Martínez et al., 2022), including attempts to establish readability metrics for legal texts (Han et al., 2024). By contrast, in the Czech Republic, the readability of legal or administrative texts has been examined only recently (Šamánková and Kubíková, 2022; Bohuslav Halfar and Bučková, 2022; Chromý et al., 2021).

As Šamánková and Kubíková (2022) mention, a quality yet clear legal text should have a clear structure, should include unambiguous instructions for the addressee and all relevant information, should be brief, legally precise and correct, and written in simple and easy-to-understand language. Tomášek (2003) and Chromá (2016) list comparable properties. As studies show (Team, 2014), the vast majority (up to 80%) of addressees (general public or experts) prefer legal or administrative texts written in a simple and clear manner. Martínez et al. (2022) show that legal texts containing features such as legal jargon of low frequency and passive voice are more difficult to comprehend, while noting that some of the features pose greater difficulties than others. In addition, the British government suggests writing in such a way that would be suitable to "*9 year old reading age*".[1]

Šamánková and Kubíková (2022) present a manual designed for authors of official administrative or legal texts, especially texts addressing the general public without extensive (or any) legal knowledge. It advises, among other things, to avoid an abundance of verbal nouns, passive voice, loanwords, and negations, and to eliminate, if possible, redundant expressions, repetitions, and abstract expressions. Similarly, Šváb (2023) advises to avoid synonyms or passive voice and to pay attention to the length of sentences, the distance between verbs and subjects, and the clear structure of the text.

In terms of regulations for the drafting of legislative texts, few sources can be found. For example, Důvodová zpráva k občanskému zákoníku (roughly translated as "Explanatory Memorandum to the Civil Code") is relatively brief while commenting

Figure 1: Front-end interface of the PONK's rule-based module

on the Civil Code in this regard.[2] It mentions that one stipulation should not contain more than two paragraphs, while one paragraph should not consist of more than two sentences (but it does not address the length of sentences). Homonyms and polysemous words are supposed to be avoided, as well as multiple different terms for an identical concept (neither of which is strictly adhered to, see Hodková (2022)). The Memorandum also mentions that loanwords are not desirable while (vaguely defined) exceptions are allowed (see Hodková (2020) for a more detailed study on this topic). It also briefly defines the usage of some expressions (in the context of the law in question).

In practice, readable legal texts differ from less readable ones in linguistic features, although linguistic readability advice is not always easy to follow (Kraus, 2025). Despite a wide range of factors contributing to (un)readability, language remains an important component according to focus groups (Veřejný ochránce práv, 2024).

3 Overview of the PONK tool

PONK is a client-oriented tool that addresses the clarity and readability of legal texts (Novotná et al., 2025).[3] It is intended for authors of legal texts. The objective of PONK is to identify phenomena that hinder readability of the text and signal them to the author. In addition to the rules presented in this paper, PONK also incorporates additional modules for readability assessment. These include (1) a lexical surprise module, which analyzes the probability of occurrence of a given token in the given context (compared to other possible tokens); e.g., a high level of lexical surprise indicates a sudden change of context, and (2) a speech acts module, which

[1] https://www.gov.uk/guidance/content-design/writing-for-gov-uk

[2] http://obcanskyzakonik.justice.cz/images/pdf/Duvodova-zprava-NOZ-konsolidovana-verze.pdf.

[3] The name itself is an acronym of "Psaní Orientované Na Klienta", approximately translated as "Client-oriented writing". The PONK tool is available at https://quest.ms.mff.cuni.cz/ponk/.

distinguishes between argumentative and normative texts and assigns to each of these two genres its typical inventory of so-called Speech Acts or Rhetorical Roles, that are based on the theory of legal writing. PONK also offers a metrics overview of the entire text, such as vocabulary, ratio and distance of verbs, or common readability metrics.[4] Figure 1 shows the front-end of the PONK's rule-based module.

PONK rules are applied to texts processed with UDPipe 2, a tool that performs tokenization, part-of-speech tagging, lemmatization, and dependency parsing according to the Universal Dependencies framework (UD),[5] and NameTag 3, a tool for named entity recognition.[6] In the Universal Dependencies framework, each token in a sentence is assigned a part-of-speech tag and represented as a node in a dependency tree.

Instead of attempting to simplify texts automatically (with the danger of altering the semantics), PONK functions as a diagnostic aid that signals segments that potentially deserve closer attention during the process of their drafting. The final decision on whether to modify the indicated passages is up to the author.

4 Implementation

The module runs as a server, and may be called via REST API.[7] We implemented the rules using the Udapi Python library.[8]

The input is supplied in the CoNNL-U format.[9] The output consists of the same CoNNL-U file with rule annotations in the MISC column, and of meta-information about the applied rules (such as rule names or user hints).

We stick to a few technical principles across all the rules. Many rules require counting (e.g. of the number of tokens in a given span). However, we found that simply counting the number of tokens may become misleading as the cohesion of constituents that tokens can combine into varies, especially in relation to grammatical agreement between nouns and their attribute(s) or determinant(s).

To illustrate this decision, we propose two sentences (A) and (B). In both sentences, the object and its governing verb (both marked by bold font) are separated by eight tokens but the respective syntactic structures are of different complexity levels.

(A) "*Stanovený* ***rozsah***$_{\textit{OBJECT-ACC}}$ *omezení*$_{\textit{GEN}}$ *má opatrovník*$_{\textit{NOM}}$ *povinnost*$_{\textit{ACC}}$ *při výkonu*$_{\textit{LOC}}$ *opatrovnictví*$_{\textit{GEN}}$ ***reflektovat*** [...]" (approx.: *The guardian has a duty to reflect the established scope of restrictions during the exercise of guardianship*. (B) "*není zřejmé,* [...] ***koho***$_{\textit{OBJECT-ACC}}$ *konkrétně předmětná*$_{\textit{NOM-ADJ}}$ *stavba*$_{\textit{NOM}}$ *svým*$_{\textit{INS-DET}}$ *závadným*$_{\textit{INS-ADJ}}$ *technickým*$_{\textit{INS-ADJ}}$ *stavem*$_{\textit{INS}}$ ***ohrožuje***." (approx.: *It is not clear [...] exactly whom the subject building endangers with its defective technical condition.*) In (A), apart from the verb "*má*" (*has*), which is part of the verbo-nominal predicate "*mít povinnost*", and the preposition "*při*" (*during*), the two tokens are separated by five nouns in four different grammatical cases (cf. the subscripts), indicating a very diverse syntactic structure. On the other hand, in (B) the two tokens are separated by one adverb and two nominal phrases, each containing a single noun preceded by one to three attributes that match the noun's grammatical case,[10] forming a less diverse, more easy-to-navigate syntactic structure.

We therefore consider adjectives and determinants to form single units with their governing tokens, which they are in morphological agreement with. Furthermore, we ignore punctuation, prepositions, and conjunctions, as they only mark syntactic relations. We call the resulting units "phrases", as they serve as a proxy for phrases with comparable perceived difficulties during parsing. The phrases are used to determine any distance, position, length, or ratio referred to in our rules (see Section 5), unless otherwise specified.

In addition, we decided to exclude citations (marked by quotation marks) from the rules' scopes, as the user will likely want to keep them intact. The module recognizes quotations by matching two quotation marks.

In this study, we call segments of texts indicated by the rules "violations", as they violate the good practices in legal texts drafting.

[4] For more details, see: https://github.com/ufal/ponk

[5] https://lindat.mff.cuni.cz/services/udpipe/, Straka (2018).

[6] https://lindat.mff.cuni.cz/services/nametag/, (Straková and Straka, 2025).

[7] The source code is available at https://github.com/ufal/ponk-linguistic-rules.

[8] https://udapi.github.io/

[9] https://universaldependencies.org/docs/format.html

[10] It should be mentioned for the sake of completeness that apart from grammatical case, this type of morphological agreement in Czech also involves grammatical number and gender.

5 Rules

In total, there are twenty rules implemented in PONK. In addition to the definition of the rules offered in this paper, the rules are also briefly described in the PONK's user manual (see Hladká et al. (2025)).

5.1 Sources of rules

One of the sources of inspiration for the rules were scientific works addressing readability. Šamánková and Kubíková (2022) and Šváb (2023) target the legal and administrative domains and the desirable drafting style within them. Sgall and Panevová (2014) refer to the "good standards" of standard Czech, although not necessarily in the context of the legal domain.

5.2 Categories of rules

We divide the rules into four categories as presented in the following list. From the perspective of traditional linguistic layers, the rules are related to syntactic or lexical phenomena.

1. Deprecated constructions and expressions (see Section 6)
2. Syntactic malconstructions rules
   (a) Position rules (see Section 7.1)
   (b) Ambiguity rules (see Section 7.2)
   (c) Cluster rules (see Section 7.3)

6 Discouraged constructions and expressions

Discouraged expressions refer to expressions that are vague or ambiguous or otherwise decrease the readability, brevity, or easy-to-navigate structure of the text. In total, eight rules belong to this category. Because most rules from this category are searching for selected phrases, exceptions for common legal terms are included in each rule if relevant.

1. ANAPHORICREFERENCES: This rule concerns vague references to anaphoric elements, which may cause uncertainty and need to examine the preceding text to identify the referent. Examples include "*Z výše uvedeného je zřejmé . . .*" (*From the above mentioned, it can be concluded . . .*) and "*S ohledem na tuto skutečnost . . .*" (*With regard to this fact . . .*).

2. ABSTRACTNOUNS: Semantically vague nouns are often used as placeholders for a more precise expression. Common examples include "*podstata*" (*nature*), "*aspekt*" (*aspect*), "*stupeň*" (*instance*), "*okolnosti*" (*circumstances*), etc. The nouns are contextually dependent, and they may occur in texts in a non-vague sense or as parts of legal terms. For example, *stupeň* may occur as part of the term "*soud prvního stupně*" (*court of first instance*).

3. CONFIRMATIONEXPRESSIONS: We define confirmation expressions as expressions that, instead of supposedly increasing certainty, paradoxically contribute to ambiguity. Examples include "*nepochybně*" (*without doubts*) and "*rozhodně*" (*surely*). On the other hand, expressions like "*jednoznačně*" (*unambiguously*) are not included because they do, in fact, increase the certainty of interpretation.

4. PASSIVEVOICE: Passive voice is grammatically correct and acceptable in Czech. However, it is considered slightly archaic and more difficult to navigate for readers, compared to active voice Šamánková and Kubíková (2022), which is recommended. That being said, passive voice is useful for sentences with unknown agent or for manipulating word order for the purposes of functional sentence perspective Firbas (2009). We therefore choose to only include passive constructions with an overt agent.

5. REDUNDANTEXPRESSIONS: Redundant expressions are expressions that have little to no function in text and can be omitted without loss or alteration of semantic content. Common examples include "*v neposlední řadě*" (*last but not least*), "*je nutné zdůraznit*" (*it is necessary to emphasize*), or "*v kontextu věci*" (*in the context of the matter*).

6. RELATIVISTICEXPRESSIONS: This rule concerns expressions that increase uncertainty. Examples include lexical units such as "*snad*" (*perhaps*, "*jaksi*" (*somehow*), "*obdobně*" (*similarly*, "*jevit*" (*to seem*), etc. We are aware that the ambiguity of some lexical units (e.g., "*velmi*", (*very*) is contextually dependent. For this reason, we focus on expressions that are inherently uncertain.

7. TOOLONGEXPRESSIONS: This rule identifies expressions that can be replaced with synonymous, but shorter equivalents to ensure the conciseness of the text. The rule searches for several specific expressions, including "*v důsledku toho*", "*v případě, že*", or "*za situace*", which can be replaced by "*proto*" (*therefore*), "*pokud*" (*if*), and "*když*" (*when*), respectively.

8. WEAKMEANINGWORDS: This rule signals lexical units (mostly verbs) that are considered se-

mantically vague and are often used as fillers in legal texts. Examples include "*zdát se*" (*to seem*) or "*ovlivnit*" (*to influence*). Users are encouraged to use more precise verbs instead. Let us note that the semantically vague verbs do not include verbs that participate in verbo-nominal predicates.[11]

7 Syntactic malconstructions rules

Suboptimal syntactic constructions, although grammatically correct, decrease the readability of the texts. Based on the nature of the constructions and the caused effect, we divide the rules into three subcategories: position rules, ambiguity rules, and cluster rules. The numeric thresholds relevant to individual rules were set after repeated experiments.

7.1 Position rules

The five position rules concern the distance between tokens with specific syntactic categories. Greater distance may make a sentence difficult to navigate, and a full understanding may require several re-reads (Šamánková and Kubíková, 2022). The position or distance of the relevant token(s) is established according to the method described in Section 4.

1. PREDSUBJDISTANCE: This rule addresses the distance between the predicate and the subject. The limit distance is 6. Violations of this rule may be triggered, among others, by adnominal clauses. In case of analytic predicate (verb forms including auxiliary verbs, modal verbs, or a copula; it is not to be confused with verbo-nominal predicate), the position of the auxiliary, copula, or the governing modal verb is taken into consideration, as they enter into agreement with the subject in Czech grammar.

2. PREDOBJDISTANCE: The limit distance between the predicate and the object is set to 6. In case of analytic predicates (see PREDSUBJDISTANCE above), we consider the position of the content token (e.g., "*byl* ***proveden***", *was* ***conducted***), as the object belongs to its valency.

3. MULTIPARTVERBS: This rule addresses the distance between the parts of multipart verbs forms (passive voice, future tense, past tense, conditional mood). Reflexive verbs are not included, because the position of the clitic "se" is fairly restricted in Czech grammar and thus mostly not pertinent for our purpose. Constructions with modal verbs are covered by the rule INFVERBDISTANCE (see below). The limit distance is 5.

4. INFVERBDISTANCE: The distance between verb and a dependent infinitive may concern two types of units: (1) constructions with modal verbs, and (2) infinitives that are objects dependent on a non-modal verb. The default limit distance is 5.

5. PREDTOOFARINCLAUSE: Although the dominant word order of Czech is SVO, it is possible (and grammatically correct) to create sentences with alternative word orders (including positioning the subject or the verbal predicate at the end of a sentence). According to the approach of functional sentence perspective (see Firbas (2009); Jasinskaja and Šimík (2023)), the non-dominant word orders are typically related to the information structure of the sentence. Due to declension, misinterpretations of such sentences are rare (although they may lead to temporary garden-path situations, Cehákova and Chromý (2023)).

As for the rule PREDTOOFARINCLAUSE, the tolerated limit position of a predicate is the 9th position from the beginning of a clause (not a whole sentence). We attempt to exclude[12] sentences with verbs at the end of a clause, since such word order can be interpreted as a likely conscious choice of putting the verb in focus (Jasinskaja and Šimík, 2023). The violations often occur in sentences with complex adjuncts.

7.2 Ambiguity rules

Ambiguity rules refer to situations in which suboptimal syntactic structures result in ambiguous constructions and decreased readability. Currently, there are two ambiguity rules.

1. DOUBLEADPOS: This rule addresses syntactic structures in which there are at least two coordinated tokens, of which the first is preceded by an overt preposition and the latter is not, leading to a possible ambiguity as for syntactic relations. An example may be "OSPOD však upozorňuji, že by se měl zabývat oznámeními otce o možné manipulaci dětí ze strany prarodičů, případně i jiných zásazích prarodičů do jeho výchovy." The absence of repetition of "*o*" leads to two potential interpretations: Either *[. . .]reports about manipulation of the children by their grandparents or about other interferences by the grandparents.* or *[. . .]reports about*

[11] For a more detailed analysis of verbo-nominal predicates in Czech, see Radimský (2017).

[12] The exclusion is currently formalized as the simultaneous occurrence of a predicate appearing with the last 3 tokens of a clause and being preceded by all its core arguments (including an infinitive).

manipulation of the children by their grandparents or by other interferences by the grandparents. If the distance between the two tokens is 4 or less, the rule is not triggered.

2. INCOMPLETECONSTRUCTION: This rule captures missing parts of multi-word constructions. If a part is omitted without the reader's knowledge, it may lead to confusion or force the reader to re-read sections of the text, as they still expect the missing part to appear later. Examples include "*jednak ... jednak*" (*both ... and*), "*buď ... nebo*" (*either ... or*), and "*zaprvé ... zadruhé*" (*firstly ... secondly*).

7.3 Cluster rules

The five rules in this subcategory concern clusters of phenomena within short spans (usually a sentence or part of a sentence) of text. For further details on how the lengths of sentences or ratios of phenomena are counted, see Section 4.

1. CASEREPETITION It is possible to encounter sequences in which many tokens share the same morphological case, but not as a result of morphological agreement. Consequently, parsing may require more cognitive load on the readers. The typical case that leads to violations of this rule is genitive, which is used for noun adjuncts in Czech.[13] The rule focuses on nouns and noun adjuncts ignores other parts of speech like adjectives or pronouns, which we consider to be easy to navigate in the text. The maximum tolerated number of nouns sharing the same case is 4. Appositions and coordinations are excluded.

2. LONGSENTENCES This rule examines the length of a sentence. The maximum length allowed is 22. The recommended modification of long sentences is to divide them into several shorter ones.

3. TOOFEWVERBS In this rule, the ratio of verbs to the length of a clause is analysed. The proportion considered as insufficient is 10% or higher. For this rule, every verbal lexeme (including modal verbs) is counted as one unit, regardless of whether the verb is complex or not.

4. TOOMANYNEGATIONS In Czech, negation concerns especially verbs, nouns, adjectives, and adverbs can be subjected to negation by the prefix *ne-*.[14] Multiple successive negations are generally discouraged (albeit grammatically correct) as they decrease readability. The rule is applied if at least 3 negations are present in a text span, and the leftmost and right-most sentences of the span must contain at least 2 negations at the same time. Legal terms containing *ne-* are excluded from the rule: "*nezletilí*" (*minors*), "*nevinný*" (*not guilty*), "*nedbalost*" (*negligence*), etc.

5. TOOMANYNOMINALCONSTRUCTIONS The proportion of nouns in a clause should not be greater than 45%. Coordinated nouns are excluded, so that enumerations do not trigger the rule. The rule also ignores named entities (conciseness is unlikely) and abbreviations typical for Czech legal texts (e.g., "*odst.*", *paragraph*).

8 Rules application on authentic texts

We tested how successfully rules violations are identified. Furthermore, we examined whether a hypothetical modification of the identified segments leads to an increase in readability. We analyzed a corpus of authentic legal texts, compiled for the purpose of this research. Then, a human annotator evaluated each identified violation.

8.1 Corpus

We established a corpus of 30 texts (a subset of the KUK corpus)[15] produced by ombudsmen, the public defenders of rights in the Czech Republic. The texts cover a period from 2007 to 2023, with the majority being published in the 2010's (18 texts). Each of the texts was assigned a document ID. The length of the individual texts varies.[16] The texts are written by different ombudsmen (although an ombudsman may author multiple documents constituting the corpus) and cover various themes, such as children's homes, disputes in medical or education domains, social security benefits, and more. The texts were not edited or modified for the purpose of our research.

8.2 Annotation

The manual annotation was performed by a linguist with experience in legal texts and terminology. The annotator controlled all identified violations and classified them as *useful*, *not useful*, or *incorrect*.

[13] In the traditional terminology of Czech grammar, we call the syntactic function "přívlastek neshodný."

[14] Tokens belonging to other parts of speech are generally not morphologically negated, with several exceptions (some pronouns, particles, etc.).

[15] https://lindat.mff.cuni.cz/repository/items/929f7d3d-4783-4d8b-8378-ccd2f6014493

[16] The longest text has 18,090 tokens (almost 110,000 characters without spaces), while the shortest text contains only 259 tokens (1,654 characters without spaces.)

The category *incorrect* indicates falsely positive violations that do not correspond to definition of the given rule. For example, in the case of the rule INCOMPLETECONSTRUCTION, there are identified spans of texts with the construction "*bud' ... nebo*", which indicates that the second part "*nebo*" was omitted from the text, while in fact the second part is present but was not recognized by the system.[17] For this reason, such violations are evaluated as *incorrect*. Another example of incorrectly signaled violations concerns the rule LONGSENTENCES, as there are cases in which the system treated two successive sentences as a single sentence.

Both *useful* and *not useful* are used for correctly identified violations (true positives). The choice between them depends on whether the alteration of the indicated segment would contribute to a higher level of readability or not. The category *not useful* thus allows for a careful consideration of the context of each violation and examination of its impact. Correct but *not useful* violations typically consist of terminological units.[18] For example, the token "*podstata*" (*nature*) is included in the rule ABSTRACTNOUNS. However, it is part of the term "*konkursní podstata*" (approx.: *bankruptcy estate*) and for this reason it is evaluated as *not useful*, as any modification would alter the legal term. *Not useful* is also applied to any identified violation within a direct citation (see Section 4).

8.3 Results

With regard to the manual annotation, we can formulate several conclusions.

First, the frequency of violations of individual rules differs according to the character of the rule. As observed in Figure 2, LONGSENTENCES tracks the highest number of identified violations, followed by TOOFEWVERBS, PREDSUBJDISTANCE, and ABSTRACTNOUNS. The frequency of these violations seems to reflect the general perception of legal texts as texts of complex and hard-to-navigate syntactic structure and a certain degree of vagueness. By contrast, violations of INCOMPLETECONSTRUCTION show the lowest absolute frequency, which is somewhat expected given the character and functions of the covered units. Perhaps curiously, violations REDUNDANTEXPRESSIONS and RELATIVISTICEXPRESSIONS also have relatively low frequencies, although we might expect them to occur more frequently in legal texts, as they both contribute to ambiguity and complexity. The low frequency may be influenced by the individual texts in the corpus or by the lexical units included in these two rules. More testing would be required to verify either hypothesis.

We can also state that the length of a document correlates with the frequency of violations (see Figure 3). As can be seen, text no. 3442 (18,090 tokens) contains a high number of violations across different rules, followed by texts no. 5648 (10,664 tokens) and no. 1962 (7,605 tokens). Observed minor deviations from this trend are connected to particularities of individual texts or influenced by the personal style of the author of the given text or by the topic of the text (especially with regard to not-yet-excluded terminological units).

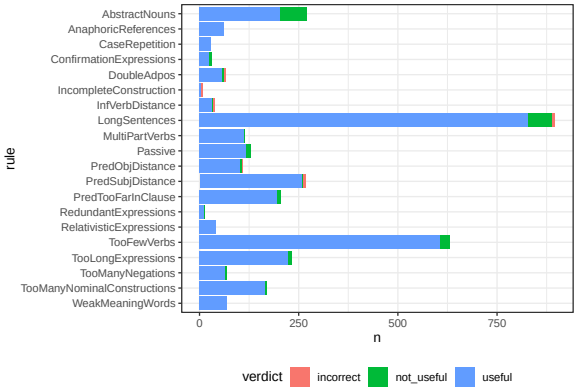

Figure 2: Overall frequencies of violations.

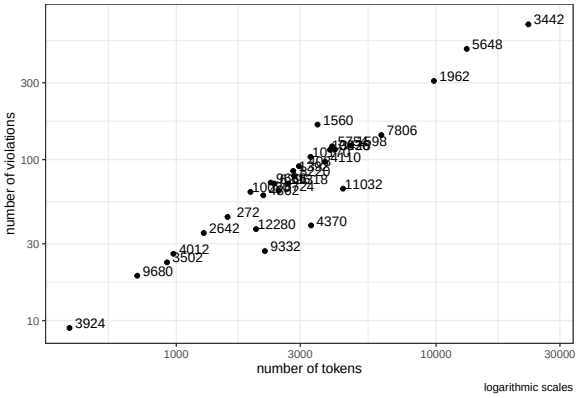

Figure 3: Number of all violations per document over document length. Document IDs are shown as labels.

[17] An example from our corpus is: "*Doporučuje se **bud'** ve VŘ podrobněji popsat pravidla pro vyřizování stížností (kdo, v jaké lhůtě, jakým způsobem), **nebo** takovou informaci vydanou ředitelem ústavu vyvěsit například na nástěnce oddělení.*" Verbatim in English: "*It is recommended to **either** describe in detail the rules for processing complaints [...], **or** post the information issued by the office's director on the office's notice board.*"

[18] Terminological units are generally excluded from the analysis, but as the exceptions are added manually, any not-yet-included term term may be indicated by the system.

In terms of the ratios of *useful*, *not useful*, and *incorrect* violations, the results are presented in Figure 4. As can be seen, *useful* indications dominate and exceed 75% in the case of all rules except INCOMPLETECONSTRUCTION. The high proportional rate of *incorrect* violations regarding this rule is due to several false positives, in which the system failed to recognize the second part of the multipart construction. Given the low absolute number of violations (7, out of which 3 are *incorrect*), we cannot identify the cause of false positives with certainty. We suspect an inaccurate text segmentation, or it may be anecdotal, but we have yet to verify this hypothesis. In the case of DOUBLEADPOS and INFVERBDISTANCE, the *incorrect* violations are caused mainly by false signaling of different parts of speech (such as verbal nouns in the case of INFVERBDISTANCE). We suspect that these false positives may be caused by the tokenization process performed by UDPipe but this remains to be verified.

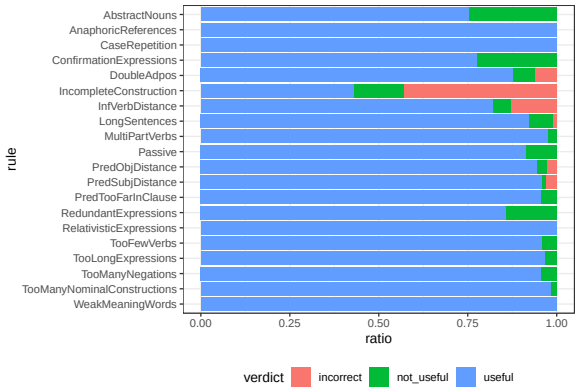

Figure 4: Ratios of useful, not useful, and incorrect violations.

The relatively high ratio of *non-useful* violations of ABSTRACTNOUNS is caused by terminological units that contain a trigger expression. These terms are to be included in the system as exceptions. In addition to terms, the token "*jasně*" (*clearly*) will become an exception to the rule CONFIRMATIONEXPRESSIONS. Similarly, we identified "*v této situaci*" (*in this situation*) as a possible exception to the rule REDUNDANTEXPRESSIONS. As for the other *not useful* instances, virtually any rule's violations can occur in direct citations (see Section 8.2 where we discussed this in detail).

9 Discussion

The results of our experiments are promising. Based on human evaluation, if the text segments indicated by our rules were modified, it might significantly increase the readability and clarity of the text and reduce ambiguity and vagueness. Thus, the PONK tool seems to provide useful feedback to users and guides them towards good practices in legal or administrative text drafting in Czech.

However, some rules still exhibit a non-trivial amount of *not useful* identified violations. This is mostly the case for the discouraged-constructions rules, as they focus on hand-picked phrases or their variations. But since meaning is generally contextually dependent and our approach does not allow us to adequately represent semantics (we only operate with lemmatization and syntactic parsing), it remains challenging to identify only phrases that truly violate the good practices while excluding "non-problematic" ones. Similarly, a considerable number of *not useful* violations is due to their terminological character or participation in direct citations. Although we tried to exclude such situations from our analysis, the results show that there is still room for improvement and further refinement.

Despite the fact that not all identified violations were evaluated as useful, we do not assume that it would significantly hinder user experience. In addition, recommendations for each violated rule and user's manual with more detailed definitions and examples are available in the PONK UI, providing explanations and a guide for potentially contentious or controversial violations.

10 Conclusions

In this paper, we presented a rule-based system, supported by a web application, that serves as an advisory tool called PONK, designed for authors of legal texts to help them increase the clarity and readability of texts. The system is adapted to Czech language and its grammar.

Following an experiment in which the system identifies phenomena non-conform to the established rules set (called "violations") in authentic Czech legal texts and human annotation of the violations, we can formulate several conclusions based on the results. The majority of the rules show high percentage (over 75%, some of the rules up to 100%) of correctly indicated violations that, if eliminated, would contribute to a more readable legal text. That being said, the experiment revealed

that some of the rules are not applied by the system as expected, and thus more testing on a larger corpus and modifications are necessary. Similarly, a thorough analysis of legal terminology that could potentially trigger the rules is required.

Despite these obstacles, we consider the system's overall performance satisfactory and positively contributing to the readability of Czech legal text in two main aspects. First, it will direct users' attention towards the parts of texts that contain characteristics of legalese or suboptimal linguistic expressions that may distract the reader. And second, providing recommendations on how to improve such segments of text might positively contribute to the readability of the users' future texts, possibly outreaching the selection of phenomena the rules are designed to capture.

Regarding future research in this area, several more rules are currently being considered for implementation. We also experimented with deterministic repair suggestions for multiple rules. However, for most rules, the applied tools lacked sufficient understanding of the linguistic structure (especially semantics and pragmatics), which would be crucial for offering reliable suggestions. The current state of the PONK tool does not include Large Language Models (LLMs) due to (1) their non-determinism and (2) privacy concerns. However, the potential positive impact of a hybrid system that includes LLMs guided by readability principles is something we are currently exploring, combining the linguistic precision of the PONK rules with the flexibility and creativity of LLMs.

Limitations

Although our rules are applied exclusively to legal texts, the vast majority of the targeted linguistic phenomena are not unique to the legal domain. Syntactic complexity, vague expressions, or structurally dense constructions can occur in virtually any Czech text, written or spoken. However, they tend to be particularly prominent in legal and administrative writing, where traditional drafting conventions often favor formal and complex formulations. The resulting misunderstandings, ambiguity, or reduced readability can directly affect the interpretation of legal texts, which underscores the importance of addressing linguistic complexity in this domain.

Our presented paper is focused on the Czech legal language and the examined phenomena are closely tied to specific grammatical and stylistic properties of Czech. For this reason, the rules cannot be transferred directly to other languages without modification. Nevertheless, we believe that the proposed rule set and the underlying methodological approach may serve as a useful starting point for similar systems developed for other languages. Some rules would likely require only minor adjustments, while others might be less applicable in languages with different grammatical structures (e.g., rules related to declension such as CASEREPETITION).

Since we focused on designing rules that provide useful feedback, we may have made some of them too strict to detect all violations of good readability practices targeted by the rules. Our evaluation design is not able to reflect on this, as it only explores the correctness or usefulness of the rules after they have been applied but not if they are applied in all relevant cases. However, we believe that the users' interaction with our module may still make them proactively notice such violations when writing other texts in the future.

Another limitation of the presented research is that it does not address legal terminology. Specialized legal terms undoubtedly contribute to the perceived opacity of legal texts, particularly for readers without legal training. Such readers may not know the precise definitions of such terms or may interpret them only through their generalized meanings in everyday language, as noted by Cornu (2005). At the same time, legal terminology constitutes an inherent part of legal discourse and therefore cannot simply be rewritten or replaced without affecting legal precision. For this reason, our present work focuses on other linguistic aspects of readability. Nevertheless, the question of accessibility of legal terms to non-expert readers remains an important direction for future research.

Our objective is not to create another normative grammar handbook, but rather to contribute to the development of tools that promote clearer and more reader-friendly legal and administrative writing. From this perspective, any step that improves the accessibility of legal language for the general public is both meaningful and necessary.

Acknowledgements

This research received financial support by the projects: "PONK - Client-Oriented Writing" (TAČR Sigma, No. TQ01000526), "Human-

centred AI for a Sustainable and Adaptive Society" (No. CZ.02.01.01/00/23_025/0008691, co-funded by the European Union and LINDAT/CLARIAH-CZ (No. LM2023062, supported by the Ministry of Education, Youth and Sports of the Czech Republic).

The work described herein has been using services provided by the LINDAT/CLARIAH-CZ Research Infrastructure (https://lindat.cz).

References

Kurt Baldinger. 1984. *Vers une sémantique moderne*. Klincksieck.

Libuše Halfarová Bohuslav Halfar and Michaela Bučková. 2022. Linguistic and legal comprehensibility of the text. *Ekonomická revue - Central European Review of Economic*.

Markéta Cehákova and Jan Chromý. 2023. Garden-path sentences and the diversity of their (mis)representations. *PLoS ONE*, 18(7).

Robert P. Charrow and Veda R. Charrow. 1979. Making legal language understandable: A psycholinguistic study of jury instructions. *Columbia Law Review*.

Marta Chromá. 2016. *Právní překlad v teorii a praxi. Nový občanský zákoník*. Karolinum.

Jan Chromý, Silvie Cinková, and Jana Šamánková. 2021. Srozumitelnost českého odborného a úředního textu — proč se jí zabývat a jak ji měřit. *Studie z aplikované lingvistiky*, 12(1):38–52.

Gérard Cornu. 2005. *Linguistique juridique, 3e édition*. Paris : Montchrestien.

Shari Seidman Diamond, Beth Murphy, and Mary R Rose. 2012. The kettleful of law in real jury deliberations: Successes, failures, and next steps. *Nw. UL Rev.*, 106:1537.

Jan Firbas. 2009. *Functional Sentence Perspective in Written and Spoken Communication*. Cambridge University Press.

Yu Han, Aaron Ceross, and Jeroen H. M. Bergmann. 2024. The Use of Readability Metrics in Legal Text: A Systematic Literature Review. *arXiv preprint*. ArXiv:2411.09497 [cs].

Alissa J. Hartig and Xiaofei Lu. 2014. Plain English and legal writing: Comparing expert and novice writers. *English for Specific Purposes*, 33:87–96.

Barbora Hladká, Silvie Cinková, Jan Černý, Vítek Eichler, Tomáš Knap, Ivan Kraus, Barbora Kubíková, Ivana Kvapilíková, Jiří Mírovský, Tereza Novotná, Tomáš Polák, Arnold Stanovský, Jana Šamánková, Michal Kuk, and Přemysl Pospíšil. 2025. Srozumitelnost českých právních a administrativních dokumentů ve výzkumu a praxi. Technical Report TR-2025-75, Prague, Czech Republic.

Kateřina Hodková. 2020. Romanisms in the Czech New Civil Code and their French Equivalents. *Acta Fakulty filozofické Západočeské univerzity v Plzni*, 12(1):61–82.

Kateřina Hodková. 2022. Les relation sémantiques au carregour des champs conceptuels du droit tchèque et du droit français. *Studia Romanistica*, 22(1).

Kateřina Hodková, Jana Pešková, Ivo Petrů, and Jan Radimský. 2021. Kontrastivní srovnávání právní terminologie na příkladech z vícejazyčné databáze právních termínů legterm. *Jazyk a kultúra*.

Katja Jasinskaja and Radek Šimík. 2023. Slavonic free word order. To appear in The Oxford guide to Slavonic languages (eds. Jan Fellerer & Neil Bermel).

Ivan Kraus. 2025. Predicting readability of czech legal writing using linguistic features.

Eric Martínez, Francis Mollica, and Edward Gibson. 2022. Poor writing, not specialized concepts, drives processing difficulty in legal language. *Cognition*, 224.

Tereza Novotná, Jan Černý, Ivan Kraus, Ivana Kvapilíková, Jiří Mírovský, Arnold Stanovský, and Barbora Hladká. 2025. PONK: Tool for Client-Oriented Legal Writing in Czech. In *JURIX 2025: The Thirty-eighth Annual Conference*, volume 416 of *Frontiers in Artificial Intelligence and Applications*, pages 330–336, Amsterdam, Netherlands. University of Turin, Italy, IOS Press.

Jan Radimský. 2017. Analytický predikát s kategoriálním slovesem. *Nový encyklopedický slovník češtiny*.

Petr Sgall and Jarmila Panevová. 2014. *Jak psát a jak nepsat česky*. Karolinum.

Milan Straka. 2018. UDPipe 2.0 Prototype at CoNLL 2018 UD Shared Task. *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*.

Jana Straková and Milan Straka. 2025. NameTag 3: A tool and a service for multilingual/multitagset NER. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 31–39, Vienna, Austria. Association for Computational Linguistics.

The GDS Team. 2014. Guest post: Clarity is king – the evidence that reveals the desperate need to re-think the way we write. *Government Digital Service*.

Michal Tomášek. 2003. *Překlad v právní praxi, 2nd edition*. Linde.

Veřejný ochránce práv. 2024. Testování srozumitelnosti. Výzkumná zpráva KVOP-17384/2024, Veřejný ochránce práv, Brno.

Jana Šamánková and Barbora Kubíková, editors. 2022. *Jak psát srozumitelné úřední texty*. Kancelář veřejného ochránce práv.

Jakub Šváb. 2023. *Jak psát, aby se to dalo číst*. Leges.