Language and Cognition (2026), 18, e16, 1–30
doi:10.1017/langcog.2026.10066

ARTICLE

Sentence-level detection of Hungarian plain language with feature-guided augmentation

István Üveges

ELTE Centre for Social Sciences, Hungary
Email: uvegesistvan898@gmail.com

(Received 28 June 2025; Revised 19 January 2026; Accepted 20 January 2026)

Abstract

In this study, we investigate Hungarian Plain Language (PL) and Simple Language (SL) with the primary objective of training a machine-learning-based sentence-level PL model that flags sentences where expert intervention may be needed during PL-oriented rewriting. The analysis uses a legal-administrative PL corpus and a news-based SL corpus, currently the only publicly available high-quality Hungarian resources for PL and SL. In low-resource settings, PL data are typically scarce, so selective data augmentation is a natural candidate for improving model performance. Our aims are threefold: (i) to provide a feature-based descriptive comparison of these Text Simplification resources; (ii) to test whether selectively chosen SL sentences can augment PL training data; and (iii) to evaluate the impact of such augmentation on sentence-level PL detection. Methodologically, we extract handcrafted linguistic features spanning surface, morphosyntactic and discourse properties. We derive a PL-likeness score from logistic-regression coefficients and use it to select SL sentences most similar to PL for augmentation, followed by supervised sentence-level PL detection with XLM-RoBERTa-large. Results show clear differences between PL and SL in sentence length, lexical diversity, syntactic depth and connective use. Selective inclusion of SL sentences yields modest gains in constrained settings, whereas indiscriminate mixing reduces precision and reliability.

Keywords: data augmentation; plain language; register analysis; simple language; text simplification

1. Introduction

Plain Language (PL) and Easy/Simple Language (SL) are clearly distinguished in literature and serve different target groups and communicative purposes. Plain Language focuses on clear, accurate and comprehensible communication, primarily in legal and administrative texts, where the aim is to ensure accessibility for a broad audience while preserving precision. By contrast, Easy Language applies stronger simplification and is intended for readers with lower literacy or limited language skills, for example, people with intellectual disabilities (Vecchiato, 2022). Thus, the

two registers are fundamental tools for tailoring communication to audiences with different goals and competencies. Both PL and SL can be viewed as Text Simplification registers, making SL a natural candidate for cross-domain augmentation. Although the international literature clearly distinguishes PL and SL as tailored language varieties (Bhatia, 2023; Leskelä et al., 2022), systematic, feature-based, corpus-driven comparisons are still lacking for Hungarian. To our knowledge, this is the first study to analyze Hungarian PL and SL as distinct registers through unified, feature-guided comparison grounded in authentic, real-world corpus data. Using the only two publicly available, high-quality Hungarian resources for PL and SL, a legal-administrative PL corpus and a news-based SL corpus, we examine their stylistic, morphosyntactic and discourse-level characteristics. This pairing reflects the current state of publicly available Hungarian resources for PL and SL. Accordingly, we note that differences in text type and specialization may introduce domain effects beyond accessibility, and we address this in our analysis. The PL corpus consists of legal and administrative documents rewritten by linguists for general audiences, whereas the SL corpus contains news texts adapted for readers with low literacy or limited proficiency. Both datasets were created for real-world use, enabling practical insights beyond experimental settings.

Our primary objective is sentence-level PL detection with a Machine Learning model that flags sentences where expert intervention may be needed during PL-oriented rewriting. Sentence-level granularity is useful in editorial workflows and is well-suited to low-resource settings where only small, high-quality PL corpora are available. Because sentences far outnumber paragraphs, it typically yields enough training instances to meet minimum sample requirements for supervised classifiers, whereas paragraph-level detection may be infeasible in small corpora. While some PL properties are realized at the discourse or text level, sentence-level instances nonetheless provide sufficient signal for models to learn the features needed for PL detection.

Methodologically, we combine a feature-based descriptive comparison with a selective augmentation experiment. First, we extract handcrafted linguistic features at the surface and morphosyntactic levels and compute sentence-level statistics on automated HuSpaCy annotations (e.g., surface metrics, morphosyntactic richness and discourse connective usage), linking these features to PL guideline concepts. Second, we derive a PL-likeness score from a logistic-regression model trained on PL data and use the score, based on top-ranked coefficients, to identify SL sentences most similar to PL. We then train and evaluate a supervised, sentence-level PL detector on PL-only data and on PL plus the selected SL instances.

The study asks three questions:

(i) Which linguistic features distinguish Hungarian PL from SL in actual use?
(ii) Can selectively chosen SL sentences improve PL detection under data scarcity?
(iii) How sensitive are PL classification results to source provenance and to the selectivity of augmentation?

Our experimental workflow consisted of seven major steps, as follows.

1. Handcrafted sentence-level surface and morphosyntactic features were extracted from HuSpaCy automatic annotations (Orosz et al., 2022) and were linked to PL guideline concepts.

https://doi.org/10.1017/langcog.2026.10066 Published online by Cambridge University Press

2. A descriptive comparison of Hungarian PL and SL was provided along these features.
3. A logistic regression model was fitted on the PL data, and a PL-likeness score was computed for SL sentences based on the top-ranked coefficients.
4. The highest-scoring SL sentences were selected for data augmentation.
5. A PL-only baseline model was trained for supervised sentence-level PL detection.
6. The detector was retrained on PL plus selected SL sentences and was evaluated against the PL-only baseline.
7. Performance differences were reported, and robustness checks were conducted, including sensitivity to the selection threshold and to source provenance.

We test whether measurable sentence-level differences between PL and SL can support supervised PL detection and whether feature-guided, low-proportion augmentation with SL sentences improves performance without degrading precision.

The remainder of the paper is organized as follows: Section 2 outlines the theoretical foundations and situates PL and SL in the Hungarian context, Section 3 introduces the data and preprocessing, including source provenance and limitations, Section 4 provides a quantitative profiling of linguistic features across the PL and SL corpora, Section 5 details the methods, that is the feature set and PL-likeness scoring as well as the supervised detection setup with augmentation experiments, Section 6 presents results and discusses implications, and Section 7 concludes and outlines future work.

2. Theoretical foundations

2.1. Conceptual and empirical perspectives on plain and simple language

Text simplification is a well-established subfield of Natural Language Processing that seeks to improve the accessibility of written texts without altering their intended meaning (Kim et al., 2024). A key conceptual distinction exists between PL and SL. In practice, however, widely used sentence-aligned simplification benchmarks, for example, work built around Simple Wikipedia (Yasseri et al., 2012) often reflect SL-style objectives, which may blur the distinction in computational settings (Stodden, 2024). Our contribution in this regard is empirical: we use corpus evidence to examine how PL and SL are realized in Hungarian. In this section, we synthesize prior work on PL and SL to derive the linguistic dimensions and features that later guide our corpus-based analysis.

2.1.1. Simple language vs plain language

SL is designed for individuals with limited literacy, non-native speakers, or those with cognitive impairments. It prioritizes basic vocabulary and simplified syntactic structures, often at the expense of semantic richness and precision (Yasseri et al., 2012). Its guiding principles include short and simple sentences, active voice, overt textual scaffolding and explicit discourse markers (Alarcon et al., 2023; Madina et al., 2023; Yaneva et al., 2016), aligning SL closely with literacy accessibility and second-language acquisition (Martínez et al., 2024).

In contrast, PL enhances comprehensibility for the average reader while maintaining legal or administrative accuracy in official settings (Fernández-Silva & Núñez Cortés, 2025; Masson & Waldron, 1994). In German linguistic literature, this contrast

https://doi.org/10.1017/langcog.2026.10066 Published online by Cambridge University Press

appears as Leichte Sprache (SL) versus Einfache Sprache (PL) (Pottmann, 2020; Radünzel, 2017).

Much of the computational simplification literature has centered on sentence-level simplification using resources with SL-like targets (e.g., Simple Wikipedia), whereas PL imposes domain-specific constraints that permit fewer transformations and require preserving normative content while improving syntactic, lexical and discourse-level accessibility (Bhatia, 2010; Plain Language Action and Information Network, 2011b; Trosborg, 1997). In legal and administrative contexts, PL revisions aim to improve clarity without altering the document's legal effect (International Organization for Standardization, 2023; Office of the Parliamentary Counsel, Drafting Techniques Group, 2014; Plain Language Action and Information Network, 2011a). By contrast, SL/Easy-to-Read permits stronger adaptations, such as re-ordering information or omitting non-essential content, to meet cognitive-accessibility needs (Inclusion Europe, 2009; Maaß, 2020; Netzwerk Leichte Sprache e.V., 2022). Because legal effect must be preserved, PL rewriting cannot materially change sentence semantics, omit information, or alter legally relevant orderings – editing is therefore more constrained than in SL, even though SL's linguistic prescriptions are often stricter.

2.1.2. From definition to features

The Plain Language Movement has evolved into a multidisciplinary research field spanning linguistics (Lutz, 2019), law (Williams, 2015) and cognitive science (Cheung, 2017). The International Plain Language Federation (2023)[1] defines a document as plain if its intended audience can find what they need, understand what they find and use that information effectively. This is codified in ISO 24495-1:2023 (International Organization for Standardization, 2023).

These definitions imply linguistic properties that support comprehension and reduce cognitive load, including shorter sentences (Matthews & Folivi, 2023), active voice (Olson & Filby, 1972) and a reduced syntactic complexity (Eslami, 2014). These tendencies accord with psycholinguistic findings on working memory capacity limitations and incremental parsing (Just & Carpenter, 1992; Van Dyke & McElree, 2006). For SL/Easy-to-Read, European and national guidance similarly articulate purposes, target audiences and drafting principles (Inclusion Europe, 2009), and peer-reviewed work frames SL and PL within register/variety perspectives and reports empirical effects and evaluations (Hennig, 2022; Leskelä et al., 2022; Vanhatalo et al., 2022).

Lexically, PL favors high-frequency, familiar, concrete vocabulary (Kimble, 2023) and discourages nominalizations (Wydick, 2005). Syntactically, complex structures such as center-embedding, multiple negation, Light Verb Constructions and over-used passives increase processing demands (M. Butt, 2010; Gibson, 1998; Liu & Cui, 2025). Maintaining canonical word order and limiting sentence length aids readers across proficiency levels (R. P. Charrow & Charrow, 1979; Pléh, 1981), consistent with broader evidence that syntactic complexity and hierarchical embedding impair comprehension (Friederici, 2011; Levy, 2008).

In SL, lexical choice is constrained toward basic, high-frequency vocabulary with explicit explanations where needed, while syntax favors short sentences, limited embedding, and canonical order; sentence splitting and overt cohesion devices are

[1] https://www.iplfederation.org/plain-language/

https://doi.org/10.1017/langcog.2026.10066 Published online by Cambridge University Press

common (Madina et al., 2023; Paetzold & Specia, 2016; Sulem et al., 2018; Yaneva et al., 2016).

Structural layout and document design can further support overall comprehensibility. The Federal Plain Language Guidelines emphasize short paragraphs, topic sentences and presenting complex information in lists or tables (Plain Language Action and Information Network, 2011a). Redundant modifiers, exceptions to exceptions and unexplained cross-references should be avoided (V. R. Charrow et al., 2007). Neurocognitive and eye-tracking research indicates that such layout principles reduce cognitive load and facilitate both global and local processing (Rayner, 1978). SL guidelines likewise prescribe overt textual scaffolding, such as descriptive headings, increased white space and supportive visuals to facilitate comprehension (Inclusion Europe, 2009; Netzwerk Leichte Sprache e.V., 2022).

Corpus-based studies show that legal and administrative texts characteristically employ passives, nominalizations and embedded clauses, which reduce readability (Bhatia, 2010; Tiersma, 1999). A comparative analysis in Hungarian (Vincze, 2018) confirms similar tendencies and variation across legal subgenres, suggesting that simplification strategies must be tailored to genre-specific conventions. In Hungary, the Miskolc Legal Corpus program (Szabó & Vinnai, 2018) reports corpus-linguistic analyses of subject and predicate structures (Balogh, 2018; Dobos, 2018), subordination and coordination (Kurtán, 2018; Sajgál, 2018) and punctuation patterns (Kurtán, 2018). Stereotypes about legalese (pervasive nominalization, function-word overload and impersonal constructions) were only partially confirmed (Vincze, 2018), while comprehension barriers were associated with more frequent departures from canonical word order (e.g., object fronting to clause-initial position) and increased distances between the verb and its dependents. In SL-oriented corpora and evaluations, simplification often increases the explicitness of discourse relations and global cohesion.

Finally, ISO 24495-1:2023 organizes PL principles around relevance of content, navigability, comprehensibility and usability. Although developed for English-language documents, its core principles have been adapted for legal and administrative domains more broadly, including in Hungary. This framing positions PL not merely as a normative ideal but as an approach aligned with cognitive constraints on language processing (Hagoort, 2019; Johnson-Laird, 1995; Kintsch, 2018).

2.2. Plain and simple language: register perspectives

PL and SL are key registers in accessibility-focused communication. Following register-oriented accounts, PL and SL are treated here as language varieties distinguished by target audiences, communicative purposes and constraints (Fuchs et al., 2023; Maaß, 2020). In this framework, PL is constrained by fidelity to institutional intent and legal/administrative accuracy, whereas SL is constrained by maximizing comprehensibility for readers with limited linguistic resources. This register perspective clarifies the functional profile of both PL and SL and guides the operationalization used later.

Functionally, PL emphasizes clarity for intended users while preserving legal/administrative effect and content; surface simplification and transparent structure are pursued without semantic loss. SL allows stronger simplification for low-literacy or cognitively diverse readers, favoring basic/high-frequency lexicon, short

https://doi.org/10.1017/langcog.2026.10066 Published online by Cambridge University Press

sentences, and explicit scaffolding and cohesion, and allowing explanatory additions or selective omission of non-essential details where appropriate. This contrast implies different error tolerances in modeling (e.g., lower tolerance for meaning shift in PL than in SL).

Differences in lexical, syntactic and discourse constraints have consequences for corpus design, annotation standards and computational modeling. Register-aware datasets and labels are needed so that features reflect register constraints rather than only genre or topic. Clarifying how PL and SL operate within and across domains is critical for aligning empirical methods with theoretical definitions and for advancing computational text simplification.

Due to the limited parallel data, most simplification research has focused on sentence-level methods, especially for low-resource, morphologically rich languages (Nisioi et al., 2017; Sulem et al., 2018; Xu et al., 2015). Sentence-level alignment can be automated with tools that match similar sentences across text variants (Alva-Manchego et al., 2020). However, recent studies highlight the limitations of sentence-only simplification, emphasizing discourse context and information structure. Sentence deletion depends on discourse factors (Zhong et al., 2020), and iterative sentence-level edits can degrade document coherence (Cripwell et al., 2023), which motivates register-aware document-level evaluation.

Although multilingual corpora like MultiSimV2 (Ryan et al., 2023) have advanced the field, Hungarian remains underrepresented. The huPWKP corpus (Prótár et al., 2023), based on English Simple Wikipedia, offers aligned data but reflects SL-style and is machine-translated into Hungarian, reducing target-language authenticity. The HunSimpleNews corpus (Prótár & Nemeskey, 2025) provides manually aligned standard and simple Hungarian news articles, focusing also on SL.

There is only one dedicated Hungarian PL corpus (Üveges, 2022). It contains aligned legal-administrative texts (the originals paired with PL rewrites by trained linguists to preserve normative content), but its modest size constrains data-intensive models. Together, these resources motivate a register-aware, feature-based comparison and sentence-level PL detection with selective SL augmentation, tested explicitly to avoid register conflation.

3. Data

Our investigation uses two Hungarian corpora: the legal–administrative Plain Language (PL) corpus introduced by Üveges (2022) and the news-based Simple/Easy Language (SL) corpus HunSimpleNews (Prótár & Nemeskey, 2025). These are, to our knowledge, the only publicly available, high-quality Hungarian resources for PL and SL. The pairing thus reflects data availability rather than a claim of domain comparability; source provenance (legal-administrative versus journalistic) is treated as a limitation throughout, and comparisons are interpreted as register-aware and descriptive. Both resources comprise documents produced for practical use, and both can be re-cast into sentence-level instances suitable for supervised model training and evaluation.

Compared with the original studies introducing the PL corpus (Üveges, 2022) and HunSimpleNews (Prótár & Nemeskey, 2025), we made preprocessing changes to ensure cross-dataset consistency. First, all texts were re-segmented at the sentence level using huSpaCy's transformer model (Orosz et al., 2022) to standardize sentence

https://doi.org/10.1017/langcog.2026.10066 Published online by Cambridge University Press

boundaries. The original PL corpus relied on a heuristic tool[2] and HunSimpleNews did not specify segmentation. Second, lead paragraphs were removed from HunSimpleNews, retaining only article bodies. For the PL corpus, section headers, tables of contents, document titles and tables were excluded, keeping only the main text bodies.

3.1. Plain language corpus: description and annotation

The PL corpus (hereinafter: Plain Language Corpus – PLC) was introduced by Üveges (2022). It comprises official texts from the National Tax and Customs Administration, Hungary (Nemzeti Adó- és Vámhivatal, NAV). These are functional informational texts intended to inform the authority's clients; the primary addressees are lay readers rather than legal professionals, and the materials are not legislative acts but citizen-facing guidance. Documents originally drafted in complex, legalistic Hungarian were rewritten by linguistic professionals to conform to institutional PL guidance, resulting in two versions per document.

Internal NAV drafting notes document the transformation principles. Although these predate ISO 24495-1, they are consistent with it (e.g., reducing passive constructions, avoiding light-verb constructions, avoiding multiply embedded clauses, preferring active voice, foregrounding reader needs and front-loading the main message). The rewriting targets clarity, navigability and audience-appropriate terminology while preserving legal/administrative effect.

The corpus spans administrative genres such as informational brochures, guidance documents and public announcements. Each source text is paired with a PL rewrite (document-level pairing). Following Üveges (2022), during sentence-level corpus preparation, sentences identical across versions were excluded. The remaining sentences were divided into sub-corpora based on presence: those found only in the PL version were labeled 'plain', and those appearing only in the original version were labeled 'non-plain'. After preprocessing, the corpus contains 4,666 sentences in the non-plain variant and 4,963 in the plain variant (Table 1).

Examples (1)–(2) are an illustrative minimal sentence pair from the PLC, where example (1) is the non-plain original, and example (2) is a plain rewrite of the same sentence.

(1) *A felugró ablakban a Mégsem funkciógombra* ***való koppintás hatására*** *az eredetileg elmentett összekapcsolás megmarad és* ***nem kerül felülírásra*** *az új összerendeléssel.*
[In the pop-up window, as a result of tapping the Cancel button, the previously saved linkage remains and is not overwritten by the new mapping.]

Table 1. Sentence and token count in the PL corpus

	Non-plain	Plain
Sentences	4,666	4,963
Tokens (lemmatized)	144,809	139,367

[2]https://github.com/mediacloud/sentence-splitter

https://doi.org/10.1017/langcog.2026.10066 Published online by Cambridge University Press

(2) *A felugró ablakban a Mégsem funkciógombra* ***koppintva*** *az eredetileg elmentett összekapcsolás megmarad és az új összerendelés* ***nem írja felül****.*
[In the pop-up window, tapping the Cancel button keeps the previously saved linkage, and the new mapping does not overwrite it.]

In this example, (1) employs a deverbal noun *koppintás* ('tapping') combined with the effect-marking postposition *hatására* ('as a result of', lit. 'on the effect of'), and it uses a periphrastic passive (*nem kerül felülírásra* – 'is not overwritten'), which increases nominalization and reduces transparency. By contrast, (2) replaces these with a verbal adverbial (*koppintva*, 'by tapping') and an active predicate (*nem írja felül*, 'does not overwrite'), yielding a more linear, sentence-level realization associated in our corpus with PL instances.

3.2. HunSimpleNews corpus: description and annotation

HunSimpleNews (HSN) is a publicly available resource for Hungarian text simplification.[3] It comprises 2,833 article pairs manually identified as true simplification matches from the PannonRTV news portal, where simplified versions were written according to Easy-to-Read principles for readers with low literacy or limited language proficiency.

After preprocessing (lead removal and huSpaCy sentence segmentation), the corpus contains 26,343 sentences in the standard language variant and 29,650 in the simplified variant (Table 2). As a journalistic resource mediated by professional writers, HSN operationalizes SL constraints (short sentences, basic lexicon, explicit scaffolding); therefore, its communicative goals and genre conventions differ from those of legal-administrative PL.

The simplifications found in the corpus are illustrated by examples (3)–(4), where (3) is the standard and (4) is the simple version of the same statement.

(3) *Nem 17 órakor fog kezdődni a kijárási tilalom, hanem 18 órától, minden nap.*
[The curfew will not start at 5 p.m. but at 6 p.m., every day.]

(4) *Keddtől a kijárási tilalom 1 órával később fog kezdődni.*
[From Tuesday, the curfew will start one hour later.]

Taken together, the corpora enable an initial, register-aware exploration of PL versus SL in Hungarian under real-world constraints. We explicitly test whether selective cross-register augmentation can support sentence-level PL detection. A full qualitative account of specialized versus journalistic style is beyond the scope of

Table 2. Sentence and token counts in the HunSimpleNews corpus

	Standard	Simplified
Sentences	26,343	29,650
Tokens (lemmatized)	496,417	344,165

[3]https://huggingface.co/datasets/ELTE-DH/HunSimpleNews

https://doi.org/10.1017/langcog.2026.10066 Published online by Cambridge University Press

this section; we flag provenance as a limitation and interpret quantitative differences accordingly.

4. Quantitative profiling of linguistic features across corpora

This section reports descriptive, sentence-level profiles of lexical, morphosyntactic and discourse-related metrics for the two corpora (PLC and HSN). The measures – covering sentence length, structural complexity and compositional patterns – outline how the datasets vary and provide a compact statistical portrait of each. Overall, the profiles jointly quantify within-register variation (PL: plain versus non-plain in official documents; SL: simple versus original in news) as well as the broader cross-register contrast between official-document PL and news-based SL.

4.1. Sentence length

One of the most salient differences between PL and SL is sentence length. As Table 3 shows, PLC sentences average 25.52 tokens and 156.54 characters, while HSN sentences are significantly shorter, averaging 15.01 tokens and 79.06 characters. Thus, PL sentences are approximately 70% longer than SL sentences in both token and character count.

This difference reflects distinct functional goals. SL texts minimize cognitive load for non-prototypical readers, such as those with cognitive or linguistic impairments. In contrast, PL texts aim to preserve legal and administrative precision while enhancing clarity for general audiences without specialized legal knowledge (P. Butt & Castle, 2006). In addition, the contrast is compounded by the baseline tendency of official documents to contain longer sentences.

4.2. Lexical diversity and repetition

Lexical diversity also reveals striking differences between PL and SL. Type–Token Ratio (TTR), the proportion of unique words (types) to total words (tokens) measures lexical richness (Baayen, 2001). Table 4 shows TTR and hapax legomena rates (words

Table 3. Base corpus statistics. (Avg. tokens = average number of tokens per sentence, Avg. characters = average number of characters per sentence, # Sentences = total number of sentences)

Corpus	Avg. tokens	Avg. characters	#Sentences
Simple language (HSN)	15.01	79.06	55,993
Plain language (PL)	25.52	156.54	9,629

Table 4. Lexical diversity and hapax legomena ratio across corpus and variant. (Tokens = total number of tokens, Unique Lemmas = total number of unique lemmas, TTR = type–token ratio, Hapax Ratio = proportion of words occurring only once – hapax legomena)

Corpus	Variant	Tokens	Unique Lemmas	TTR	Hapax Ratio
HSN	Simple	344,165	16,245	0.0472	0.0205
HSN	Standard	496,417	31,517	0.0635	0.0320
PL	Non-plain	125,565	9,110	0.0726	0.0310
PL	Plain	120,123	8,913	0.0742	0.0315

https://doi.org/10.1017/langcog.2026.10066 Published online by Cambridge University Press

occurring only once), calculated from lemmatized text (inflected forms reduced to their base lemma).

The PL corpus exhibits slightly higher lexical diversity: non-plain TTR is 0.0726, plain TTR is 0.0742. In contrast, SH corpus TTR is 0.0635 for standard and only 0.0472 for simple variants. Hapax legomena ratios are stable in PL (3.1% both variants) but lower in HSN/simple (2.0%) compared to HSN/standard (3.2%).

4.3. Part-of-speech distributions

Differences in Part-of-Speech (POS) distributions reveal structural divergences between PL and SL. As Tables 5 and 6 show, nouns (NOUN) dominate PL: they account for 30.1% of tokens in non-plain and 28.0% in plain variants, exceeding HSN noun ratios (22.2% standard, 26.0% simple).

This noun overrepresentation reflects the nominalization-heavy style typical of legal-administrative texts (Mattila, 2016), indicating a conceptual rather than action-oriented mode. In contrast, HSN corpus shows higher verb usage (~10.2%), usually associated with clear, event-based communication. Adjectives (ADJ) are also more frequent in PL (17.6% / 16.4%) than SL (10.2%). SL places greater emphasis on

Table 5. The top 10 POS categories in the **HunSimpleNews corpus (HSN)**, expressed as percentages. (NOUN = Nouns, ADJ = Adjectives, PUNCT = Punctuation, DET = Determiners, VERB = Verbs, PROPN = Proper Nouns, ADV = Adverbs, CCONJ = Coordinating Conjunctions, PRON = Pronouns, ADP = Adpositions)

POS	Standard (%)	Simple (%)
NOUN	22.2	26.0
PUNCT	15.8	13.4
DET	12.6	14.0
VERB	10.3	10.2
ADJ	10.0	10.2
ADV	7.0	6.7
PROPN	6.6	6.4
CCONJ	5.2	3.9
PRON	3.6	2.0
NUM	2.6	3.2

Table 6. The top 10 POS categories in the **Plain Language corpus (PLC)**, expressed as percentages. (NOUN = Nouns, ADJ = Adjectives, PUNCT = Punctuation, DET = Determiners, VERB = Verbs, PROPN = Proper Nouns, ADV = Adverbs, CCONJ = Coordinating Conjunctions, PRON = Pronouns, ADP = Adpositions)

POS	Non-plain (%)	Plain (%)
NOUN	30.1	28.0
ADJ	17.6	16.4
PUNCT	13.1	14.2
DET	12.2	12.9
VERB	6.3	7.1
PROPN	5.0	4.9
ADV	4.1	4.6
CCONJ	3.7	4.2
PRON	2.8	2.6
ADP	1.8	1.6

https://doi.org/10.1017/langcog.2026.10066 Published online by Cambridge University Press

determiners (DET) and pronouns (PRON); DET ratio in SL/simple reaches 14.0% versus 12.6% in SL/standard. Conversely, PRON usage declines from 3.6% in SL/standard to 2.0% in SL/simple, suggesting a preference for explicit noun repetition over coreference chains for easier reference resolution. Proper nouns (PROPN) also vary: SL news articles include more Named Entities (6.4% simple, 6.6% standard) than the PLC's more abstract, institutional content (4.9% plain, 5.0% non-plain).

4.4. Morphological variety and type richness

Morphological features were analyzed at the sentence level using two metrics: the average number of distinct morphological types per sentence and morphological TTR (type–token ratio), capturing variation in inflectional forms, case markers, verb tenses and agreement features (Table 7).

Morphological TTR is notably higher in HSN. Although HSN/simple texts have fewer morphological types per sentence (7.38) than PL/plain texts (11.48), their relative diversity (TTR = 0.675) exceeds that of PL/plain (0.553) and PL/non-plain (0.543). Thus, SL texts, despite syntactic simplicity, show surprising morphological variation within short sentences.

This reflects morphologically rich but structurally simple constructions, such as varied verb inflections, case-marked noun phrases and explicit adpositional expressions, enhancing referential clarity and accessibility. SL seems to prioritize clarity through morphological explicitness (e.g., pronouns, spatial adpositions) rather than syntactic embedding.

4.5. Syntactic complexity and embedding

Syntactic complexity was assessed using two metrics from huSpaCy dependency-parsed annotations:

- the dependency tree depth per sentence,
- the ratio of tokens labeled as embedded clause dependents (*csubj, ccomp, xcomp, acl*).

Tree depth was computed by mapping token indices to syntactic heads and measuring the longest path to the root per sentence. The embedded clause ratio was calculated as the proportion of such tokens relative to total sentence tokens. Sentence-level values were aggregated by corpus and variant to yield average syntactic depth, embedded clause ratio and sentence counts.

Table 7. Morphological type richness across corpus and variant. (Avg. types/sentence = average number of unique word forms per sentence, Morph. TTR = morphological type–token ratio)

Corpus	Variant	Avg. types/sentence	Morph. TTR
PL	Non-plain	12.41	0.543
PL	Plain	11.48	0.553
HSN	Standard	9.99	0.590
HSN	Simple	7.38	0.675

https://doi.org/10.1017/langcog.2026.10066 Published online by Cambridge University Press

Table 8. Syntactic complexity measures by corpus and variant. (Avg. Tree Depth = average depth of the syntactic parse tree per sentence, Embedded Clause Ratio = proportion of clauses embedded within sentences)

Corpus	Variant	Avg. Tree Depth	Embedded Clause Ratio
PL	Non-plain	4.99	0.0212
PL	Plain	4.62	0.0215
HSN	Standard	3.74	0.0216
HSN	Simple	3.01	0.0170

Table 8 summarizes the results, confirming that PLC texts are syntactically more complex than HSN. PL/non-plain has an average parse depth of 4.99, with a slight reduction in PL/plain (4.62). HSN/standard averages 3.74, while HSN/simple is flatter at 3.01.

Embedded-clause proportions are broadly stable across registers (~2.1%), with only HSN/simple showing a lower value (0.0170). Consistent with this, PL complexity likely reflects other sources (e.g., internal NP modifiers, participial constructions, legal-adverbial phrases, such as '*within 8 days from receipt of the notification*') rather than more frequent subordination.

The minimal depth reduction from PL/non-plain to PL/plain suggests simplification focuses on clarity (improved word order, reduced redundant modifiers, enhanced thematic salience) rather than flattening clause structure. Such restructuring maintains syntactic depth while improving linearization and prosodic clarity. By contrast, SL actively minimizes embedding. HSN/simple exhibits shallow parse trees and favors linear, one- or two-clause structures with low configurational complexity, aligning with its short sentences and high morphological TTR.

In summary, syntactic complexity differentiates between PL and SL. PL preserves hierarchical structure even under simplification, while SL reduces it to minimize cognitive processing demands.

4.6. Discourse connectives and cohesion strategies

Discourse connectives reveal finer-grained cohesion and coherence differences across registers. We categorized these connectives into five types:

- **contrastive** (e.g., *but, although, however*),
- **explanatory** (e.g., *because*, *since, therefore*),
- **amplifying** (e.g., *moreover, in addition*),
- **temporal** (e.g., *when, after, then*), and
- **exemplifying** (e.g., *for example*),

as these markers guide readers in navigating logical structures and inferring discourse relations.

Table 9 shows average connective counts per sentence (in the lemmatized texts). The PL corpus shows moderate usage across categories, higher than SH/simple but lower than HSN/standard. For example, contrastive connectives occur at 0.058 per sentence in PL/non-plain and 0.050 in PL/plain, while HSN/simple has a much lower rate (0.012). HSN/standard employs them more frequently (0.090).

https://doi.org/10.1017/langcog.2026.10066 Published online by Cambridge University Press

Table 9. Average number of discourse connectives per sentence across corpora and variants

Corpus	Variant	Contr.	Expl.	Amp.	Temp.	Illustr.
PL	Non-plain	0.058	0.051	0.029	0.09	0.021
PL	Plain	0.050	0.064	0.017	0.008	0.041
HSN	Standard	0.090	0.061	0.009	0.051	0.008
HSN	Simple	0.012	0.024	0.000	0.024	0.027

Note: The columns show the mean frequency of contrastive (Contr.), explanatory (Expl.), amplifying (Amp.), temporal (Temp.) and illustrative (Illustr.) markers per sentence.

Notably, explanatory connectives are more frequent in PL/plain than in PL/non-plain (0.064 versus 0.051), reflecting a strategy that favors explicit marking of logical relations rather than structural flattening. Explanatory connectives (e.g., 'because', 'namely', 'therefore') guide readers through causal, conditional or deductive relations, aligning with PL's goal of enhancing clarity. In contrast, HSN/simple shows substantial connective reduction. Amplifying, temporal and contrastive markers occur minimally, often below 0.03 per sentence.

Table 10 shows the percentage of sentences containing at least one connective per type. In PLC, about 5% of sentences contain contrastive connectives, while in HSN/simple, this is just above 1%.

In summary, PL sentences increase the visibility of logical structure through explanatory connectives, while SL sentences minimize connective use, favoring structural simplicity and reducing inferential demands.

4.7. Readability and register functionality

Despite their serious limitations (Benjamin, 2012; Crossley et al., 2008), readability metrics provide an additional lens to contrast. Table 11 presents the results of a Flesch-style readability index adapted for Hungarian, based on average sentence length and syllables per word. Higher scores (closer to 100) indicate greater ease of comprehension.

HSN/simple achieves a high average Flesch score of 79.93, categorizing it as 'very easy to read' and HSN/standard also scores relatively high (67.79).

In contrast, PL texts show lower general readability. The non-plain variant averages 34.67, with the plain version improving slightly to 41.21. Although sentence length and syllable density decrease (2.41–2.35), PL texts remain in the 'difficult' band. This can reflect both PL's functional constraints (clarifying legally complex content without sacrificing interpretative precision) and register differences between corpora (news versus official texts).

Table 10. Percentage of sentences containing discourse connectives across corpora

Corpus	Variant	Contr.	Expl.	Amp.	Temp.	Illustr.
PL	Original	5.62%	4.91%	2.76%	0.84%	2.04%
PL	Plain	4.88%	6.21%	1.55%	0.75%	3.77%
HSN	Standard	8.68%	5.91%	0.91%	4.93%	0.75%
HSN	Simple	1.15%	2.39%	0.02%	2.43%	2.74%

Note: The columns show the proportion of sentences including contrastive (Contr.), explanatory (Expl.), amplifying (Amp.), temporal (Temp.) and illustrative (Illustr.) markers.

https://doi.org/10.1017/langcog.2026.10066 Published online by Cambridge University Press

Table 11. Flesch readability scores and surface text characteristics across corpus and variant

Corpus	Variant	Avg.syll./word	Avg.sentence length	Flesch score
PL	Non-plain	2.41	26.91	34.67
PL	Plain	2.35	24.20	41.21
HSN	Standard	2.00	18.84	67.79
HSN	Simple	1.92	11.61	79.93

5. Methods

5.1. Feature extraction and selection

Feature selection builds on a targeted literature review and the corpus characteristics described earlier (Sections 2–4). We aim to capture surface, morphosyntactic and discourse-related properties characteristic of PL and SL, aligning feature engineering with prior theoretical distinctions with what our corpora make observable. Crucially, the feature set is designed not only to separate PL and SL but also to identify SL sentences that are proximate to PL because PL training data are scarce and augmentation with linguistically compatible SL instances may improve sentence-level PL detection without conflating registers.

To capture structural and stylistic characteristics of PL and SL sentences, we compute a broad set of linguistic features derived from huSpaCy's transformer model annotations, supplemented by custom lexical and discourse heuristics. Where relevant, features operationalize PL principles as discussed in Section 2 (e.g., shorter sentences, controlled syntax, transparent cohesion) and SL constraints (e.g., basic vocabulary, explicit scaffolding). The feature set covers the following dimensions:

Surface features

- **n_tokens**: Total number of tokens (words and punctuation) in a sentence.
- **n_chars**: Total number of characters in the sentence.
- **n_syllables**: Estimated total number of syllables (proxy for phonotactic difficulty/processing load).
- **mean_token_length**: Average word length (characters), indicating lexical complexity.
- **ttr**: Type–token ratio, proportion of unique words to total words (lexical diversity).
- **n_repeats**: Number of lemma repetitions (referential consistency/redundancy).
- **flesch**: Adapted Hungarian Flesch readability score (Flesch, 1948).

Morphosyntactic and syntactic features

- **depth**: Maximum depth of the dependency tree (hierarchical complexity).
- **n_subordinate_clauses**: Count of embedded clauses (*advcl, ccomp, acl dependencies*).
- **embedded_clause_ratio**: Proportion of tokens in embedded clauses.
- **n_conjunctions**: Coordinating conjunctions or conjunct relations.
- **subj_type**: Grammatical category of the main subject (e.g., PRON, NOUN).
- **n_function_words**: Number of function words (pronouns, determiners, auxiliaries).
- **lexical_density**: Proportion of content words (nouns, verbs, adjectives, adverbs) relative to total tokens (information load).

https://doi.org/10.1017/langcog.2026.10066 Published online by Cambridge University Press

- **n_topical_nouns**: Number of domain-specific (legal or administrative) nouns, representing thematic density.
- **n_modal_verbs**: Number of modal verbs (e.g., *must, may, should*).
- **n_nominalizations**: Count of deverbal noun forms, indicating abstract or formal expression level.

Discourse and referential features

- **n_anaphora**: Number of anaphoric pronouns (e.g., *this, that, they*; POS=PRON).
- **n_lists**: Number of enumerative structures (typically after a colon).
- **starts_with_connective**: Binary indicator of sentence-initial connectives.

Connective features (by rhetorical function)
For each connective type, three variables were computed:

- **n_X**: raw frequency
- **ratio_X**: relative frequency (normalized by sentence length)
- **sent_X**: binary presence flag

Here, X stands for the following connective categories:

- **contrastive**: e.g., *but*, although, however
- **explanatory**: e.g., because, since, therefore
- **amplifying**: e.g., moreover, in addition
- **temporal**: e.g., when, after, then
- **illustrative**: e.g., for example

This feature set supports interpretable machine learning models and quantitative corpus analyses, enabling precise measurement of PL–SL stylistic and structural differences. The same features underpin a PL-likeness score (Section 5.2) used to select SL candidates for augmentation. This targets data efficiency under PL scarcity and tests whether proximity-based augmentation helps detection while avoiding naïve register mixing.

5.2. Classification models and training setup

With the feature space established, we next investigated how these characteristics enable automated differentiation between PL and SL. This section describes the classification framework, including the choice of model architecture, training setup and evaluation protocols. By aligning model design with the linguistic features introduced earlier, we ensured that classifiers captured the structural and stylistic dimensions of PL in a transparent way.

5.2.1. Linguistic predictors of plain language rewriting

Logistic regression estimates statistical associations between input features and a binary target class, making it a valuable tool for interpretability in text classification (Hastie et al., 2009; James et al., 2021). Positive coefficients indicate features more common in one class (e.g., PL), while negative coefficients indicate the opposing

https://doi.org/10.1017/langcog.2026.10066 Published online by Cambridge University Press

class (e.g., non-PL). We use this coefficient profile operationally to define a PL-likeness score that ranks HSN sentences by their proximity to the PLC 'plain' sub-corpus. Together with the analogous HSN model (Section 5.2.2) and the comparison in Section 5.2.3, this enables the motivated cross-register augmentation strategy evaluated later.

We trained a logistic regression model on the PLC (Section 3.1). The model achieved F1 = 0.58 (accuracy: 0.58). Despite moderate predictive performance, the coefficient profile is useful for the interpretation of feature associations.

According to the results (Figure 1), PL sentences are associated with:

- **Higher type-token ratio** (ttr: $\beta = 0.55$).
- **More illustrative markers and lists** (n_illustrative: $\beta = 0.49$, sent_illustrative: $\beta = 0.13$, n_lists: $\beta = 0.009$).
- **More modal verbs** (n_modal_verbs: $\beta = 0.26$).
- **Higher counts of explanatory connectives** (n_explanatory: $\beta = 0.12$, sent_explanatory: $\beta = 0.08$, ratio_explanatory: $\beta = 0.08$).
- **More coordinating conjunctions** (n_conjunctions: $\beta = 0.08$).

In contrast, negative associations include:

- **Lower lexical density** (lexical_density: $\beta = -1.23$).
- **Fewer anaphora and embedded clauses** (n_anaphora: $\beta = -0.27$, embedded_clause_ratio: $\beta = -0.26$).
- **Reduced intensification and contrastive discourse markers** (sent_amplifying: $\beta = -0.26$, n_contrastive: $\beta = -0.25$).
- **Fewer intensifiers, abstract markers and nominalizations** (n_amplifying: $\beta = -0.19$, n_nominalizations: $\beta = -0.06$).

We note that ttr is length-sensitive; sentence-level results should be read with this caveat.

5.2.2. *Linguistic predictors of simple language*

We trained an analogous logistic-regression model to distinguish simple versus standard sentences in HSN. Figure 2 shows the distinct coefficient profile for SL. The strongest positive coefficients are:

- **lexical_density** ($\beta = 1.41$),
- **n_lists** ($\beta = 0.943$),
- **sent_illustrative** ($\beta = 0.897$) and **n_illustrative** ($\beta = 0.895$),
- **n_modal_verbs** ($\beta = 0.331$) and
- **ttr** ($\beta = 0.212$).

Negative coefficients are observed for:

- **n_amplifying** ($\beta = -1.456$) and **sent_amplifying** ($\beta = -1.450$),
- **n_contrastive** ($\beta = -0.689$) and **sent_contrastive** ($\beta = -0.658$),
- **n_explanatory** ($\beta = -0.261$),
- **n_topical_nouns** ($\beta = -0.817$) and
- **n_subordinate_clauses** ($\beta = -0.357$).

https://doi.org/10.1017/langcog.2026.10066 Published online by Cambridge University Press

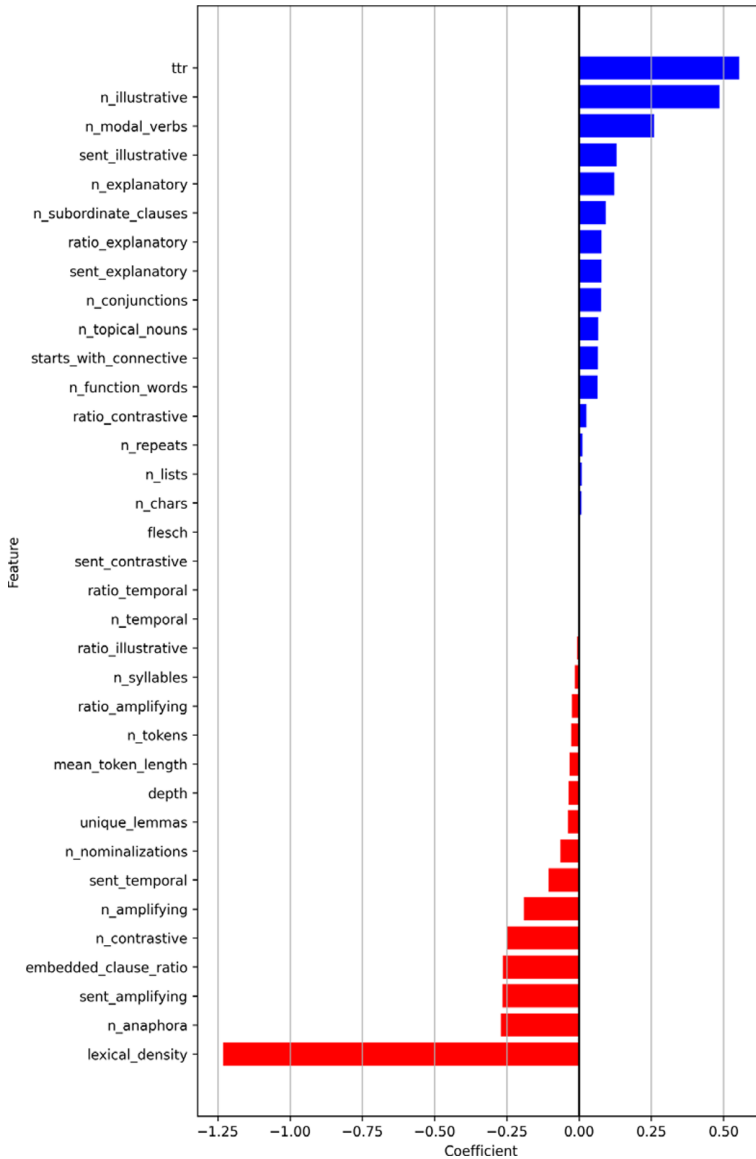

Figure 1. Coefficients from logistic regression trained on PL sentences only. Positive values (blue) indicate greater likelihood of 'plain' classification, whereas negative values (red) indicate greater likelihood of 'non-plain' classification.

5.2.3. Comparing plain and simple language registers

Building on the two models, we compared signs and magnitudes of selected coefficients (see Table 12). We observe overlaps (e.g., some positive association of ttr in both PLC and HSN) and divergences (e.g., lexical_density negatively associated with PLC 'plain' but positively with HSN 'simple'; fewer embedded clauses in both datasets but stronger list/illustrative associations in HSN). The ttr observation should be interpreted

https://doi.org/10.1017/langcog.2026.10066 Published online by Cambridge University Press

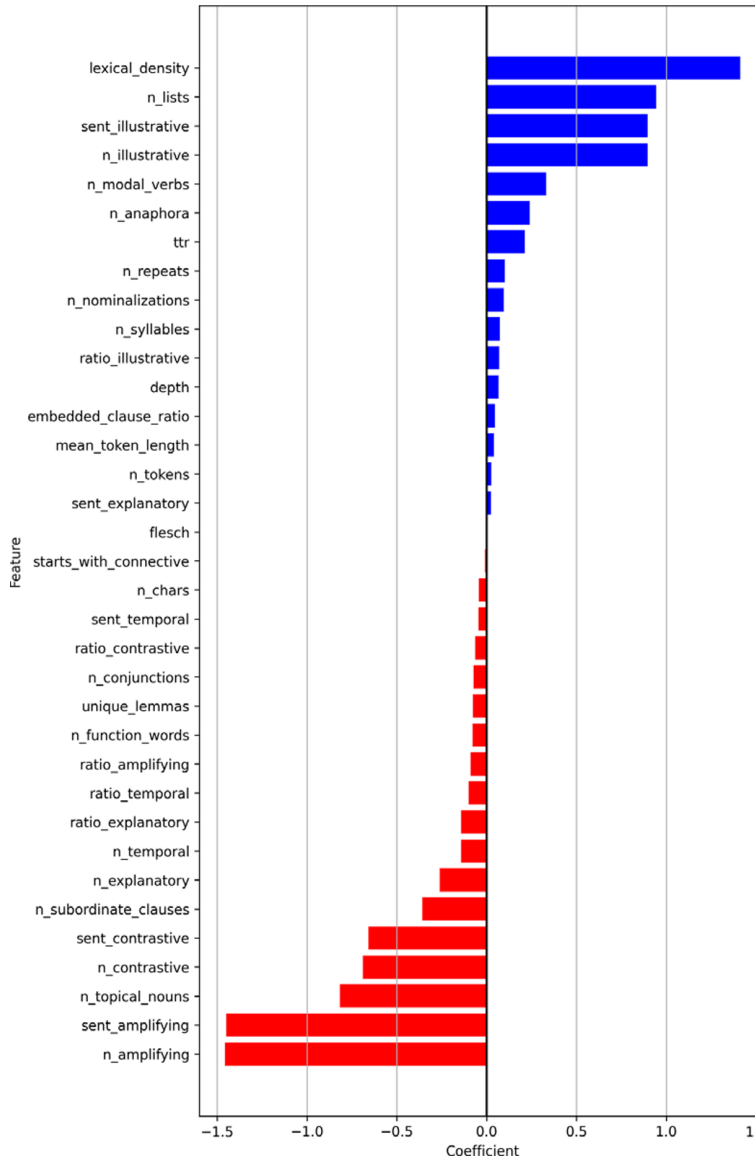

Figure 2. Coefficients from logistic regression trained on simple versus standard HSN sentences. Positive values (blue) indicate associations with the 'simple' label, whereas negative values (red) indicate associations with the 'standard' label.

cautiously because ttr is sensitive to sentence length. Regarding 'n_topical_nouns' versus lexical_density, the former indexes domain-specific terminology, whereas the latter is the ratio of content to function words; these metrics capture different dimensions and can move in opposite directions.

This coefficient profile later underpins a PL-likeness score to rank HSN sentences by proximity to PLC 'plain' sentences. From a modeling perspective, direct cross-

https://doi.org/10.1017/langcog.2026.10066 Published online by Cambridge University Press

Table 12. Comparison of logistic regression coefficients for selected linguistic features

Feature	PL	HSN	Interpretation
lexical_density	−1.23	1.41	HSN favors higher lexical density, whereas PL favors reducing the proportion of dense lexical items.
ttr	0.55	0.21	Lexical diversity is promoted in both registers, but more strongly in PL.
n_illustrative	0.49	0.895	The use of examples is a key simplification strategy for both registers.
n_modal_verbs	0.26	0.33	Explicit modal expressions (obligation, possibility) are frequently present in both registers.
n_amplifying	−0.19	−1.456	Intensifiers are avoided in both registers.
n_contrastive	−0.25	−0.689	Contrastive markers are reduced in both registers, with a stronger trend in HSN.
n_lists	≈0	0.943	Enumerations cannot be associated with PL, but are central in HSN.
n_topical_nouns	≈0	−0.817	HSN actively reduces institutional or technical vocabulary, a trend not observed in PL.
n_subordinate_clauses	≈0	−0.357	HSN systematically avoids complex hierarchical syntax. PL shows no strong trend for this feature.
n_explanatory	0.12	−0.261	Explanatory markers are retained in PL, while HSN tends to omit them.
n_anaphora	−0.27	0.24	PL reduces anaphoric pronouns, whereas HSN uses more of them.

Note: Positive values indicate association with the plain or simple label, respectively.

register transfer is likely limited. A model trained on PLC may underrepresent patterns prominent in HSN and vice versa. Accordingly, augmentation should avoid naive mixing; instead, selection is guided by proximity in the feature space (see next section).

5.2.4. Scoring-based selection of SL sentences for PL augmentation

To augment PLC with HSN instances, we implemented a coefficient-weighted scoring to identify simple sentences most similar to PLC 'plain' along the selected features. The aim is to test whether proximity-based augmentation benefits PL detection under scarcity. HSN simple sentences were ranked by a PL-likeness score computed as a weighted sum of standardized features (top-k features by absolute PLC coefficients, for details, see Appendix A):

$$\text{Score}_{\text{i}} = \sum_{j=1}^{n} w_j \times z_{ij}$$

Here, Score_i is the PL-likeness score of sentence i, n is the number of selected features, w_j is the coefficient of feature j (derived from the PL classification model), and z_{ij} is the z-score normalized value of feature j in sentence i.

This approach allowed us to isolate high-confidence PL-like instances from HSN in a controlled, interpretable manner. To ensure a proportional contribution of features to the score, we standardized them using z-score normalization (Izenman, 2008), as features are measured on different scales (e.g., number of tokens versus lexical ratios). The z-score was computed in the standard way as follows:

https://doi.org/10.1017/langcog.2026.10066 Published online by Cambridge University Press

$$z_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j}$$

Here, x_{ij} is the raw value of feature j for sentence i, and μ_j and σ_j are the mean and standard deviation of feature j computed over all sentences under consideration.

Standardization was applied to HSN only, since PLC sentences were not scored. Candidates were sorted by Score_i and the top-k% were selected ($k \in \{5, 10, 25, 100\}$). This procedure is interpretable and controllable; however, it does not guarantee gains and is evaluated empirically in Section 6.

5.3. Data augmentation techniques

As a first step, we established a baseline for PL classification and then augmented it with selected instances from the HSN corpus. Evaluation uses the same held-out PLC subset across conditions, with only training/validation composition varying.

5.3.1. Augmentation methodology

Two settings were compared:

- **PL-only model** (trained exclusively on PLC), and.
- **PLC + HSN-k%**: (PLC plus top-k% HSN simple sentences by PL-likeness).

All comparisons are on the same, fixed PLC test set.

5.3.2. Creation of augmented training sets

The original (PLC) training set contained 4,963 'plain' sentences. Four augmented sets were created by adding HSN simple sentences at 5%, 10%, 25% and 100% of the plain-class size (cf. Table 13). Only positively scored candidates were considered; in practice, this did not constrain the larger settings. All augmented sets were shuffled before training. The non-plain class remained unchanged.

5.4. Implementation details

We used XLM-RoBERTa-large (Conneau et al., 2020) as the backbone for sentence-level classification. In a recent comparative evaluation of Transformer-based models for PL classification in Hungarian (Anonymous, 2025), XLM-RoBERTa-large achieved the highest performance among all tested architectures, including the first Hungarian BERT model, huBERT (Nemeskey et al., 2021), XLM-RoBERTa-base and

Table 13. Baseline and augmented training/validation compositions

Model	Train PL	Train HSN	Val PL	Val HSN
PL–Only	3970	0	993	0
PL+HSN–5%	3970	199	993	50
PL+HSN–10%	3970	397	993	99
PL+HSN–25%	3970	993	993	248
PL+HSN–100%	3970	3970	993	993

Note: PLC-only uses only PLC instances; PLC + HSN-100% adds an equal number of HSN sentences to the corpus.

https://doi.org/10.1017/langcog.2026.10066 Published online by Cambridge University Press

proprietary LLMs such as Gemini 1.0 Pro (Google, 2023) and GPT-4o-mini (OpenAI, 2024). Its advantage was particularly evident in cross-domain evaluation and low-resource settings, where other models tended to overfit or underperform. We attribute this result to XLM-R's robust subword tokenization, strong multilingual pretraining (including substantial Hungarian coverage) and deep representational capacity, which enable it to capture register-sensitive characteristics beyond surface-level syntax or token frequency.

6. Results and discussion

6.1. Evaluation of the baseline model

As a baseline, we fine-tuned the XLM-RoBERTa-large model using only the sentence pairs labeled as changed in the Plain Language corpus (cf. Section 3.1). The model was trained as a binary classifier, distinguishing plain (label 1) from non-plain (label 0) sentences. The evaluation was performed on a held-out test set comprising 963 stratified examples.

Table 14 summarizes the performance metrics. The model achieved an overall accuracy of 69%, with an F1 score of 0.73 for the plain class and 0.63 for the non-plain. Precision and recall were reasonably balanced, although the classifier tended to favor plain labels, as shown in the confusion matrix in Figure 3.

The results suggest that the model captures important stylistic and structural features of PL sentences but struggles with ambiguous or borderline cases, particularly in identifying non-plain constructions. These limitations motivate the exploration of augmentation strategies, which are discussed in the next section.

6.2. Evaluating the effectiveness of SL-based augmentation for PL classification

This section interprets the experimental results on augmenting the PLC/plain class with systematically selected HSN sentences. The central goal was to assess whether HSN sentences with high PL-likeness can improve classifier performance for legal–administrative texts.

The measured PL-likeness scores (Figure 4) for HSN/simple sentences show that most instances cluster near zero due to z-score normalization. However, a subset falls within the top 5%–10%, suggesting that these sentences share characteristics with PL, including simplified syntax, accessible vocabulary and lower structural complexity, making them promising augmentation candidates.

However, augmentation benefits were not consistently monotonic. As shown in Figure 5, modest augmentation (5%) improved all metrics except plain class

Table 14. Classification metrics (precision, recall, F1) for the PL-only model, evaluated on the held-out PL test set (non-plain versus plain instances)

Class	Precision	Recall	F1-score	Support
Non-plain (0)	0.73	0.56	0.63	459
Plain (1)	0.67	0.81	0.73	504
Accuracy		0.69		963
Macro avg	0.70	0.68	0.68	
Weighted avg	0.70	0.69	0.69	

https://doi.org/10.1017/langcog.2026.10066 Published online by Cambridge University Press

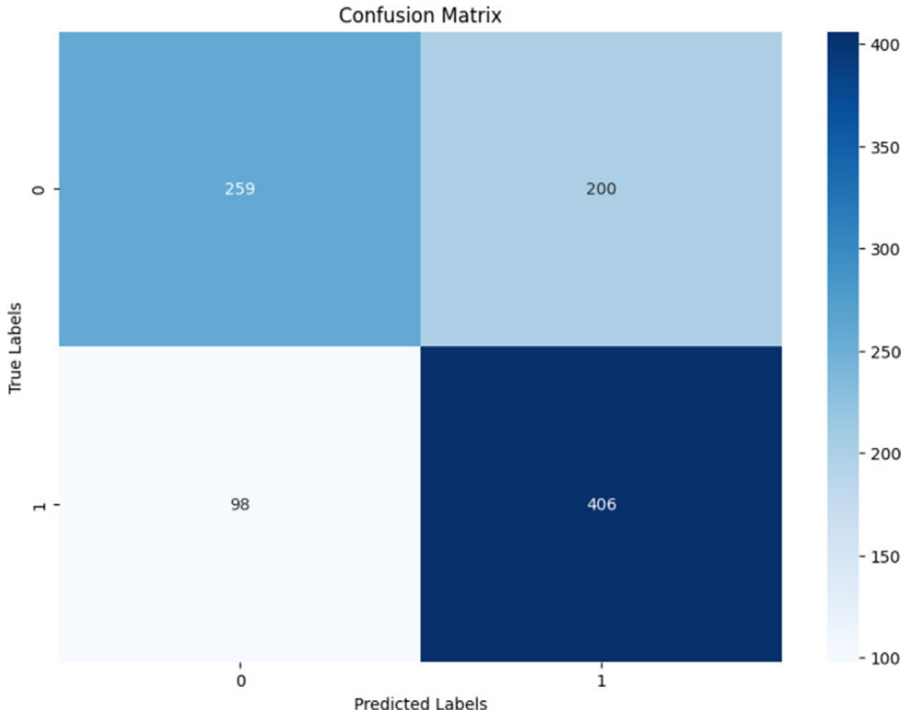

Figure 3. Confusion matrix for the PL-only model. Label 1 = plain; label 0 = non-plain.

precision, which remained unchanged from the baseline. At 10%, precision increased slightly but recall and F1 declined. This suggests that the model became more conservative in assigning the plain label (fewer false positives), but missed more true positives, as indicated by lower recall and F1. Overall performance (macro and weighted averages, Table 15) remained stable between 5% and 10%, indicating augmentation offers limited benefits beyond a narrow range.

When the augmentation rate increases to 25%, performance gains plateau. At 100%, when all HSN sentences with PL-likeness scores are added as plain examples, model discrimination deteriorates. Precision and recall for the non-PL class drop, while PL class recall rises sharply, indicating overgeneralization as the model increasingly predicts PL labels for structurally or semantically inappropriate inputs.

A critical concern is class imbalance. High augmentation rates can cause the training data to be dominated by plain-labeled sentences, as all HSN examples are assigned to this label. This shift may affect optimization dynamics, with the loss function overly influenced by the dominant class. As a result, the model becomes prone to plain class predictions, reducing sensitivity to the complex syntax and formal terminology of non-plain texts. However, balanced performance remains achievable with appropriate regularization, weighting or resampling strategies when sufficient training data are available (Buda et al., 2018).

Interestingly, overall recall improves with more augmentation (reaching 0.84 for full HSN inclusion) but at the cost of precision, which drops to 0.67, indicating more false positives for the plain class. The 5% and 25% augmentation levels yield similar

https://doi.org/10.1017/langcog.2026.10066 Published online by Cambridge University Press

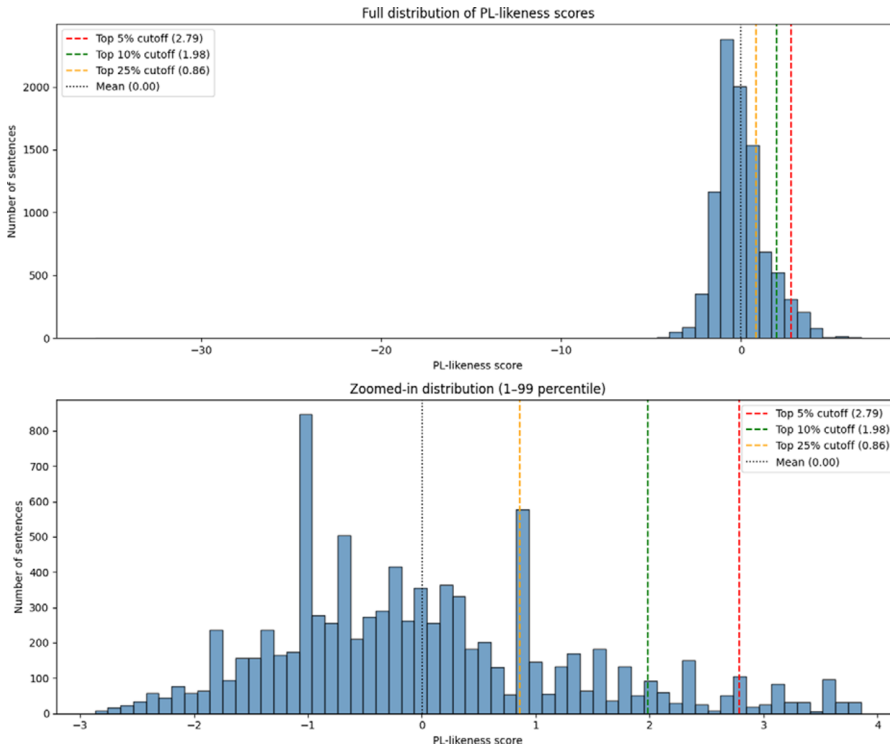

Figure 4. Distribution of PL-likeness scores for HSN/simple sentences. The top chart shows the full range, the bottom chart zooms into the 1st to 99th percentiles. Vertical lines indicate cutoffs for the top 5%, 10% and 25% of scores.

accuracy (0.70) and F1-scores (0.74 and 0.70), differing mainly in recall and precision balance. This reflects a classic trade-off: increased permissiveness captures more true positives but misclassifies more non-plain inputs.

Notably, while including all HSN sentences yields the highest recall and F1-score (0.75), this improvement warrants caution. It may result from overfitting to superficially simple but functionally misaligned patterns rather than genuine domain alignment with legal-administrative plainness. Thus, the performance boost masks degraded semantic fidelity.

This raises a second concern: the functional and register-level fidelity of the non-plain class. While augmenting it with standard HSN sentences could correct class imbalance, this strategy neglects the linguistic specificity of non-plain data. Standard HSN sentences, though not plain, are less complex than legal–administrative texts. Adding them creates a heterogeneous training signal that may impair the model's ability to learn bureaucratic and legal–administrative language characteristics defining both PLC classes. Consequently, the classifier risks learning an impoverished decision boundary. Instead of distinguishing legally normative, citizen-facing plainness from formal institutional legalese, it may only differentiate based on syntactic complexity, confounding generalization across registers and reducing utility where legal clarity is critical.

https://doi.org/10.1017/langcog.2026.10066 Published online by Cambridge University Press

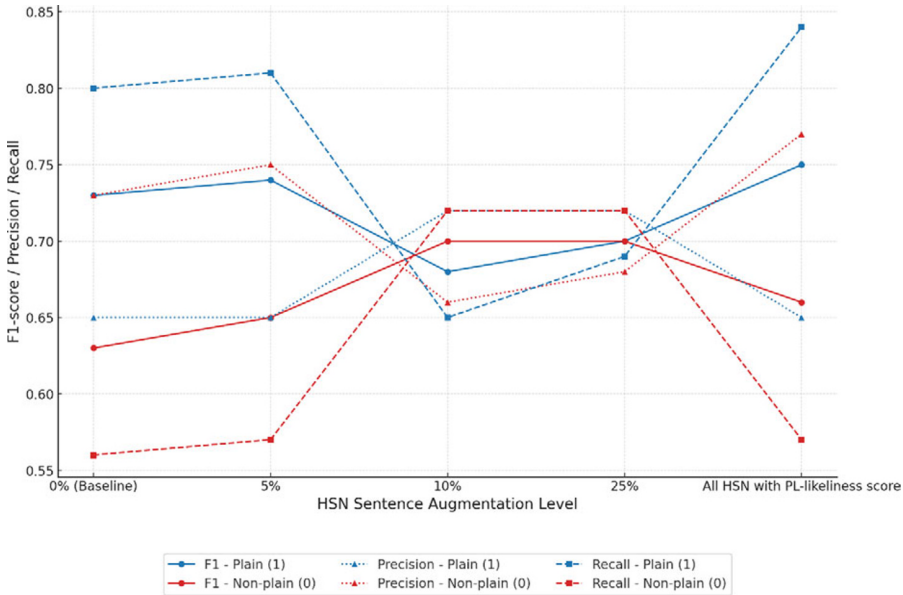

Figure 5. Effect of PL-likeness-based augmentation with HSN data on classification performance across F1-score, precision and recall for both classes.

Table 15. Classification results across augmentation levels

Model	Class	Precision	Recall	F1-score
Baseline	Non-PL (0)	0.73	0.56	0.63
	PL (1)	0.67	0.81	0.73
	Macro avg	0.70	0.68	0.68
	Weighted avg	0.70	0.69	0.69
PL+HSN–5%	Non-PL (0)	0.75	0.57	0.65
	PL (1)	0.67	0.82	0.74
	Macro avg	0.71	0.70	0.69
	Weighted avg	0.71	0.70	0.69
PL+HSN–10%	Non-PL (0)	0.66	0.74	0.70
	PL (1)	0.72	0.64	0.68
	Macro avg	0.69	0.69	0.69
	Weighted avg	0.69	0.69	0.69
PL+HSN–25%	Non-PL (0)	0.68	0.73	0.70
	PL (1)	0.72	0.67	0.70
	Macro avg	0.70	0.70	0.70
	Weighted avg	0.70	0.70	0.70
PL+HSN–100%	Non-PL (0)	0.77	0.57	0.66
	PL (1)	0.67	0.84	0.75
	Macro avg	0.72	0.70	0.70
	Weighted avg	0.72	0.71	0.70

Overall, these findings highlight that augmentation success depends not just on syntactic similarity but also on pragmatic and register alignment. PL is not merely structural simplification but a functional mode of legal communication. Augmenting PL with non-legal but structurally similar examples risks introducing noise. Similarly,

https://doi.org/10.1017/langcog.2026.10066 Published online by Cambridge University Press

augmenting the non-plain class with general language fails to reinforce features defining bureaucratic or juridical complexity.

In sum, SL-based augmentation in PL contexts is viable only under carefully constrained, linguistically informed conditions. The top 5%–10% of HSN sentences with high PL-likeness scores offer measurable improvements and are relatively safe. Beyond this, gains are illusory, and representational fidelity of both classes is compromised.

7. Conclusion

Our corpus-based analysis indicates that Hungarian Plain Language (PL) and Simple Language (SL) are related but distinct as instantiated in the two available resources we study. At the sentence level, we observe consistent differences in length, lexical diversity, syntactic depth, part-of-speech distributions and discourse connective usage. In our data, PL sentences are longer and structurally deeper on average, whereas SL sentences are shorter and achieve higher scores on a Hungarian Flesch-style index. PL/plain sentences use explanatory connectives more often than PL/non-plain, while SL/simple reduces connective use overall. At the same time, some tendencies converge across registers when comparing each simplified variant to its own original: both PL/plain and SL/simple show fewer embedded clauses than their respective counterparts. It should be noted, however, that these statements summarize observed distributions in the two corpora rather than effects outside them.

Within this empirical setting, selective inclusion of SL sentences offers limited support for sentence-level PL detection. Adding the top 5%–10% of HSN/simple sentences ranked by PL-likeness yields small, measurable gains. Beyond this range (e.g., 25%), improvements plateau, and at full inclusion model discrimination deteriorates, with increased overprediction of the plain label. These outcomes are sensitive to the augmentation threshold and reflect interactions with class composition and source provenance.

Taken together, the results support a register-aware approach to corpus design and model training for PL detection. Feature-guided selection can identify SL instances that are structurally compatible with PL and can be used in low proportions under data scarcity, but gains remain modest and depend on alignment with the target register.

Future work should expand legal-administrative PL resources and examine document-level context to test whether sentence-level tendencies persist when discourse organization is modeled directly. Given that data scarcity persists, it is also worth exploring alternative augmentation strategies, such as Easy Data Augmentation, Large Language Model-based synthetic example generation, or machine translation and multilingual models. While these are less linguistically motivated, they could exploit language models' capacity for controllable paraphrastic variation, lexical and syntactic perturbations, and class-conditioned sampling if register alignment and source provenance are monitored and evaluation remains strictly corpus-based.

References

Alarcon, R., Moreno, L., & Martínez, P. (2023). EASIER corpus: A lexical simplification resource for people with cognitive impairments. *PLoS One*, 18(4), e0283622. https://doi.org/10.1371/journal.pone.0283622.

Alva-Manchego, F., Martin, L., Bordes, A., Scarton, C., Sagot, B., & Specia, L. (2020). ASSET: A dataset for tuning and evaluation of sentence simplification models with multiple rewriting transformations. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 4668–4679). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.424.

https://doi.org/10.1017/langcog.2026.10066 Published online by Cambridge University Press

Anonymous. (2025). *Comparative evaluation of transformer-based models for hungarian plain language classification.*

Baayen, R. H. (2001). *Word frequency distributions* (Vol. 18). Springer Netherlands. https://doi.org/10.1007/978-94-010-0844-0.

Balogh, D. (2018). Alanyi szerkezetek a magyar jogi nyelvben. In M. Szabó & E. Vinnai (Eds.), *A törvény szavai: Az OTKA-112172 kutatási zárókonferencia anyaga Miskolc, ME – MAB, 2018. Május 25* (pp. 71–94). Bíbor kiadó.

Benjamin, R. G. (2012). Reconstructing readability: Recent developments and recommendations in the analysis of text difficulty. *Educational Psychology Review*, 24(1), 63–88. https://doi.org/10.1007/s10648-011-9181-8.

Bhatia, V. K. (2010). Interdiscursivity in professional communication. *Discourse & Communication*, 4(1), 32–50. https://doi.org/10.1177/1750481309351208.

Bhatia, V. K. (2023). Legal genres in interdiscursive contexts. In A. Wagner & A. Matulewska (Eds.), *Research handbook on jurilinguistics* (pp. 159–178). Edward Elgar Publishing. https://doi.org/10.4337/9781802207248.00019.

Buda, M., Maki, A., & Mazurowski, M. A. (2018). A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 106, 249–259. https://doi.org/10.1016/j.neunet.2018.07.011.

Butt, M. (2010). The light verb jungle: Still hacking away. In M. Amberber, B. Baker, & M. Harvey (Eds.), *Complex predicates* (1st ed., pp. 48–78). Cambridge University Press. https://doi.org/10.1017/CBO9780511712234.004.

Butt, P., & Castle, R. (2006). *Modern legal drafting: A guide to using clearer language second edition* (2nd ed.). Cambridge University Press. https://doi.org/10.1017/CBO9781139168533.

Charrow, R. P., & Charrow, V. R. (1979). Making legal language understandable: A psycholinguistic study of jury instructions. *Columbia Law Review*, 79(7), 1306. https://doi.org/10.2307/1121842.

Charrow, V. R., Erhardt, M. K., & Charrow, R. P. (2007). *Clear and effective legal writing*. Aspen Law & Business.

Cheung, I. W. (2017). Plain language to minimize cognitive load: A social justice perspective. *IEEE Transactions on Professional Communication*, 60(4), 448–457. https://doi.org/10.1109/TPC.2017.2759639.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 8440–8451). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.747.

Cripwell, L., Legrand, J., & Gardent, C. (2023). Document-level planning for text simplification. In *Proceedings of the 17th conference of the European chapter of the association for computational linguistics* (pp. 993–1006). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.eacl-main.70.

Crossley, S. A., Greenfield, J., & McNamara, D. S. (2008). Assessing text readability using cognitively based indices. *TESOL Quarterly*, 42(3), 475–493. https://doi.org/10.1002/j.1545-7249.2008.tb00142.x.

Dobos, C. (2018). Állítmányi szerkezetek a magyar jogi nyelvben. In M. Szabó & E. Vinnai (Eds.), *A törvény szavai: Az OTKA-112172 kutatási zárókonferencia anyaga Miskolc, ME – MAB, 2018. Május 25* (pp. 37–69). Bíbor kiadó.

Eslami, H. (2014). The effect of syntactic simplicity and complexity on the readability of the text. *Journal of Language Teaching and Research*, 5(5), 1185–1191. https://doi.org/10.4304/jltr.5.5.1185-1191.

Fernández-Silva, S., & Núñez Cortés, J. A. (2025). Exploring the effect of plain terminology on processing and comprehension of administrative texts in Spanish: A self-paced reading experiment. *International Journal of Applied Linguistics*, 35(2), 659–671. https://doi.org/10.1111/ijal.12650.

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology*, 32(3), 221–233. https://doi.org/10.1037/h0057532.

Friederici, A. D. (2011). The brain basis of language processing: From structure to function. *Physiological Reviews*, 91(4), 1357–1392. https://doi.org/10.1152/physrev.00006.2011.

Fuchs, J., Bock, B. M. & Pappert, S. (2023). Leichte Sprache, Einfache Sprache, verständliche Sprache. In *Mit Beiträgen von Pirkko Friederike Dresing, Mathilde Hennig und Cordula Meißner*. Tübingen: Narr. *Zeitschrift Für Angewandte Linguistik*, 2023(79), 262–267. https://doi.org/10.1515/zfal-2023-2014

Gibson, E. (1998). Linguistic complexity: Locality of syntactic dependencies. *Cognition*, 68(1), 1–76. https://doi.org/10.1016/S0010-0277(98)00034-1.

Google. (2023). *Gemini 1.0 Pro*. https://console.cloud.google.com/vertex-ai/publishers/google/model-garden/gemini-pro

Hagoort, P. (2019). The neurobiology of language beyond single-word processing. *Science*, 366(6461), 55–58. https://doi.org/10.1126/science.aax0289.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning*. Springer New York. https://doi.org/10.1007/978-0-387-84858-7.

Hennig, M. (2022). ETR German within the system of language variation. *Nordic Journal of Linguistics*, 45(2), 214–231. https://doi.org/10.1017/S0332586522000129.

Inclusion Europe (Ed.). (2009). *Schulungen für Lehrerinnen und Lehrer: Wie man anderen Menschen beibringt, Texte in leichter Sprache zu schreiben; [entwickelt im Rahmen des Projektes Pathways – Wege zur Erwachsenenbildung für Menschen mit Lernschwierigkeiten]*. Inclusion Europe.

International Organization for Standardization. (2023). *Plain language—part 1: Governing principles and guidelines* (No. ISO/DIS 24495–1). https://www.iso.org/obp/ui/#iso:std:iso:24495:-1:dis:ed-1:v1:en

International Plain Language Federation. (2023). What is plain language? https://www.iplfederation.org/plain-language/

Izenman, A. J. (2008). *Modern multivariate statistical techniques*. Springer New York. https://doi.org/10.1007/978-0-387-78189-1.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An introduction to statistical learning: With applications in R*. Springer US. https://doi.org/10.1007/978-1-0716-1418-1.

Johnson-Laird, P. N. (1995). *Mental models: Towards a cognitive science of language, inference, and consciousness*. Harvard University Press.

Just, M. A., & Carpenter, P. A. (1992). A capacity theory of comprehension: Individual differences in working memory. *Psychological Review*, 99(1), 122–149. https://doi.org/10.1037/0033-295X.99.1.122.

Kim, J., Leijnen, S., & Beinborn, L. (2024). Considering human interaction and variability in automatic text simplification. In *Proceedings of the third workshop on text simplification, accessibility and readability (TSAR 2024)* (pp. 52–60). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.tsar-1.6.

Kimble, J. (2023). *Writing for dollars, writing to please: The case for plain language in business, government, and law* (2nd ed.). Carolina Academic Press.

Kintsch, W. (2018). Revisiting the construction—Integration model of text comprehension and its implications for instruction. In D. E. Alvermann, N. J. Unrau, M. Sailors, & R. B. Ruddell (Eds.), *Theoretical models and processes of literacy* (7th ed., pp. 178–203). Routledge. https://doi.org/10.4324/9781315110592-12.

Kurtán, Z. (2018). Központozás a magyar jogi nyelvben. In M. Szabó & E. Vinnai (Eds.), *A törvény szavai: Az OTKA-112172 kutatási zárókonferencia anyaga Miskolc, ME – MAB, 2018. Május 25* (pp. 153–179). Bíbor kiadó.

Leskelä, L., Mustajoki, A., & Piehl, A. (2022). Easy and plain languages as special cases of linguistic tailoring and standard language varieties. *Nordic Journal of Linguistics*, 45(2), 194–213. https://doi.org/10.1017/S0332586522000142.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, 106(3), 1126–1177. https://doi.org/10.1016/j.cognition.2007.05.006.

Liu, X., & Cui, Y. (2025). Eye tracking technology for examining cognitive processes in education: A systematic review. *Computers & Education*, 229, 105263. https://doi.org/10.1016/j.compedu.2025.105263.

Lutz, B. (2019, September). Plain language and applied linguistics: An integrated view on technical communication [Conference presentation]. PLAIN 2019, Oslo, Norway. https://doi.org/10.13140/RG.2.2.31317.78565.

Maaß, C. (2020). *Easy language – Plain language – Easy language plus: Balancing comprehensibility and acceptability*. Frank & Timme. https://doi.org/10.26530/20.500.12657/42089.

Madina, M., Gonzalez-Dios, I., & Siegel, M. (2023). Easy-to-read in Germany: A survey on its current state and available resources (no. arXiv:2306.03189). *arXiv*. https://doi.org/10.48550/arXiv.2306.03189.

Martínez, P., Moreno, L., & Ramos, A. (2024). Exploring large language models to generate easy to read content (no. arXiv:2407.20046). *arXiv*. https://doi.org/10.48550/arXiv.2407.20046.

Masson, M. E. J., & Waldron, M. A. (1994). Comprehension of legal contracts by non-experts: Effectiveness of plain language redrafting. *Applied Cognitive Psychology*, 8(1), 67–85. https://doi.org/10.1002/acp.2350080107.

Matthews, N., & Folivi, F. (2023). Omit needless words: Sentence length perception. *PLoS One*, 18(2), e0282146. https://doi.org/10.1371/journal.pone.0282146.

Mattila, H. E. S. (2016). *Comparative legal linguistics*. Routledge. https://doi.org/10.4324/9781315573106.

Nemeskey, D. M., Váradi, T., & Sass, B. (2021). Hungarian BERT models for natural language processing. In *Proceedings of the 17th conference on hungarian computational linguistics* (pp. 3–14). University of Szeged, Faculty of Scienceand Informatics, Institute of Informatics.

Netzwerk Leichte Sprache e.V. (2022). *Die Regeln für Leichte Sprache*. Netzwerk Leichte Sprache e.V. https://www.netzwerk-leichte-sprache.de/fileadmin/content/documents/regeln/Regelwerk_NLS_Neuauflage-2022.pdf

Nisioi, S., Štajner, S., Ponzetto, S. P., & Dinu, L. P. (2017). Exploring neural text simplification models. In *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 2: Short papers)* (pp. 85–91). Association for Computational Linguistics. https://doi.org/10.18653/v1/P17-2014.

Office of the Parliamentary Counsel, Drafting Techniques Group. (2014). *Drafting guidance* (Nos. DEP2014–0422). Cabinet Office. London: Office of the Parliamentary Counsel. https://data.parliament.uk/DepositedPapers/Files/DEP2014-0422/guidancebook_20_March.pdf

Olson, D. R., & Filby, N. (1972). On the comprehension of active and passive sentences. *Cognitive Psychology*, 3(3), 361–381. https://doi.org/10.1016/0010-0285(72)90013-8.

OpenAI. (2024). GPT-4o system card.

Orosz, G., Szántó, Z., Berkecz, P., Szabó, G., & Farkas, R. (2022). HuSpaCy: An industrial-strength Hungarian natural language processing toolkit (version 2). *arXiv*. https://doi.org/10.48550/ARXIV.2201.01956.

Paetzold, G., & Specia, L. (2016). SemEval 2016 task 11: Complex word identification. In *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)* (pp 560–569). Association for Computational Linguistics. https://doi.org/10.18653/v1/S16-1085.

Plain Language Action and Information Network. (2011a). Federal plain language guidelines. https://www.plainlanguage.gov/media/FederalPLGuidelines.pdf

Plain Language Action and Information Network. (2011b). What is plain language? https://www.plainlanguage.gov/about/definitions/

Pléh, C. (1981). The role of word order in the sentence interpretation of Hungarian children. *Folia Linguistica*, 15(3–4). https://doi.org/10.1515/flin.1981.15.3-4.331.

Pottmann, D. M. (2020). „Leichte Sprache and Einfache Sprache" – German plain language and teaching DaF German as a foreign language. *Studia Linguistica*, 38, 81–94. https://doi.org/10.19195/0137-1169.38.6

Prótár, N., & Nemeskey, D. M. (2025). HunSimpleNews: Az első autentikus magyar nyelvű szövegekből álló szövegegyszerűsítési korpusz. *XXI. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY*, 2025, 197–218.

Prótár, N., Nemeskey, D. M., Digital Humanities, D., & Eötvös Loránd University, Hungary (2023). huPWKP: A Hungarian text simplification corpus. In *Proceedings of the conference recent advances in natural language processing – Large language models for natural language Processings* (pp. 898–907). University of Szeged, Faculty of Science andInformatics, Institute of Informatics. https://doi.org/10.26615/978-954-452-092-2_097.

Radünzel, C. (2017). Leichte Sprache: Eine linguistische Betrachtung eines neuen sprachlichen Phänomens auf der Grundlage polnischer und deutscher Beispieltexte. *Zeitschrift für Slawistik*, 62(1), 48–95. https://doi.org/10.1515/slaw-2017-0003.

Rayner, K. (1978). Eye movements in reading and information processing. *Psychological Bulletin*, 85(3), 618–660. https://doi.org/10.1037/0033-2909.85.3.618.

Ryan, M., Naous, T., & Xu, W. (2023). Revisiting non-English text simplification: A unified multilingual benchmark. In *Proceedings of the 61st annual meeting of the Association for Computational Linguistics (volume 1: Long papers)* (pp. 4898–4927). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.acl-long.269

Sajgál, M. (2018). Alárendelés a magyar jogi nyelvben. In M. Szabó & E. Vinnai (Eds.), *A törvény szavai: Az OTKA-112172 kutatási zárókonferencia anyaga Miskolc, ME – MAB, 2018. Május 25* (pp. 123–151). Bíbor Kiadó.

Stodden, R. (2024). EASSE-DE & EASSE-multi: Easier automatic sentence simplification evaluation for German & multiple languages. In *Proceedings of the third workshop on text simplification, accessibility and readability (TSAR 2024)* (pp. 107–116). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.tsar-1.11

Sulem, E., Abend, O., & Rappoport, A. (2018). Simple and effective text simplification using semantic and neural methods. In *Proceedings of the 56th annual meeting of the association for computational linguistics*

https://doi.org/10.1017/langcog.2026.10066 Published online by Cambridge University Press

(volume 1: Long papers) (pp. 162–173). Association for Computational Linguistics. https://doi.org/10.18653/v1/P18-1016

Szabó, M., & Vinnai, E. (2018). *A törvény szavai*. Bíbor kiadó.

Tiersma, P. M. (1999). *Legal language*. University of Chicago Press.

Trosborg, A. (1997). *Rhetorical strategies in legal language: Discourse analysis of statutes and contracts*. Narr.

Üveges, I. (2022). Comprehensibility and automation: Plain language in the era of digitalization. *TalTech Journal of European Studies*, 12(2), 64–86. https://doi.org/10.2478/bjes-2022-0012.

Van Dyke, J. A., & McElree, B. (2006). Retrieval interference in sentence comprehension. *Journal of Memory and Language*, 55(2), 157–166. https://doi.org/10.1016/j.jml.2006.03.007.

Vanhatalo, U., Lindholm, C., & Onikki-Rantajääskö, T. (2022). Introduction: Easy language research. *Nordic Journal of Linguistics*, 45(2), 163–166. https://doi.org/10.1017/S0332586522000130.

Vecchiato, S. (2022). Clear, easy, plain, and simple as keywords for text simplification. *Frontiers in Artificial Intelligence*, 5, 1042258. https://doi.org/10.3389/frai.2022.1042258.

Vincze, V. (2018). A Miskolc Jogi Korpusz nyelvi jellemzői. In M. Szabó & E. Vinnai (Eds.), *A törvény szavai: Az OTKA-112172 kutatási zárókonferencia anyaga Miskolc, ME – MAB, 2018. Május 25* (pp. 9–36). Bíbor kiadó.

Williams, C. (2015). Changing with the times: The evolution of plain language in the legal sphere. *Revista Alicantina de Estudios Ingleses*, 28, 183. https://doi.org/10.14198/raei.2015.28.10.

Wydick, R. C. (2005). *Plain English for lawyers* (5th ed.). Carolina Academic Press.

Xu, W., Callison-Burch, C., & Napoles, C. (2015). Problems in current text simplification research: New data can help. *Transactions of the Association for Computational Linguistics*, 3, 283–297. https://doi.org/10.1162/tacl_a_00139.

Yaneva, V., Temnikova, I., Mitkov, R., Calzolari, N., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., & Piperidis, S. (2016). Evaluating the readability of text simplification output for readers with cognitive disabilities. In *Proceedings of the tenth international conference on language resources and evaluation ({LREC}'16)* (pp. 293–299). European Language Resources Association(ELRA).

Yasseri, T., Kornai, A., & Kertész, J. (2012). A practical approach to language complexity: A Wikipedia case study. *PLoS One*, 7(11), e48386. https://doi.org/10.1371/journal.pone.0048386.

Zhong, Y., Jiang, C., Xu, W., & Li, J. J. (2020). Discourse level factors for sentence deletion in text simplification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05), 9709–9716. https://doi.org/10.1609/aaai.v34i05.6520.

Appendix A. Feature Weights Used in HSN Sentence Ranking

To identify HSN sentences suitable for augmenting the PL classification model, we ranked all HSN items using a scoring function based on the ten most influential linguistic features from the logistic regression model trained on changed PL sentence pairs. These features were selected based on the absolute value of their coefficients, indicating their predictive strength. Each HSN sentence received a composite compatibility score computed as a weighted sum of its normalized (z-scored) feature values, using the PL model's coefficients as weights.

Table A1 lists the features used in this scoring procedure along with their respective weights.

Table A1. Top 10 linguistic features and coefficients from the PL (logistic regression) model used for scoring HSN sentences

Feature	Description	Weight (β)
lexical_density	Proportion of content words	−1.231
ttr	Type-to-token ratio	+0.554
n_explanatory	Number of illustrative connectives	+0.486
n_anaphora	Number of anaphoric pronouns	−0.270
sent_amplifying	Presence of intensifying connectives	−0.265
embedded_clause_ratio	Ratio of tokens in embedded clauses	−0.264
n_modal_verbs	Number of modal verbs	+0.259
n_contrastive	Number of contrastive connectives	−0.251
n_amplifying	Number of intensifying connectives	−0.190
sent_explantory	Presence of illustrative connectives	+0.131

Note: Positive weights indicate alignment with plain language characteristics; negative weights suggest divergence.

https://doi.org/10.1017/langcog.2026.10066 Published online by Cambridge University Press

Each feature was standardized across the HSN dataset using z-score normalization:

$$z_i = \frac{x_i - \mu_j}{\sigma_j}$$

where x_i is the raw value of feature i for a given sentence, μ_j is the mean and σ_j is the standard deviation of feature i across all HSN sentences.

The final sentence score was computed as follows:

$$\text{score} = \sum_{i=1}^{10} \beta_i z_i$$

where β_i denotes the logistic regression coefficient for feature i in the PL model. Sentences with higher scores are considered more stylistically compatible with the plain register and were prioritized during data augmentation.

Cite this article: Üveges, I. (2026). Sentence-level detection of Hungarian plain language with feature-guided augmentation, *Language and Cognition*, 18, e16, 1–30. https://doi.org/10.1017/langcog.2026.10066