IMAGE: A MAP OF THE STARS OF THE ORION CONSTELLATION

JournalPreview

London Journal of Research in Computer Science & Technology

Volume 25 | Issue 4 | Compilation 1.0

Great Britain Journals Press

WWW.JOURNALSPRESS.COM

Journal Content
In this Issue

Great Britain Journals Press

i. Journal introduction and copyrights
ii. Featured blogs and online content
iii. Journal content
iv. Editorial Board Members

1. Design of a Real-Time, Multilingual, Emotion-Aware Cyberbullying Detection System using Multi-Teacher Knowledge Distillation and Explainable AI. ***1-16***
2. The Weakest Link in Internet Privacy: Security and Compliance Risks in Third-Party Vendor Data Handling. ***17-23***
3. Self-Service Analytics 2.0: AI-Powered Dashboard Generation with Human-in-the Loop Feedback Architecture. ***25-29***
4. The Relationship between Consciousness and Linguistic Data to Formalize. ***31-53***

v. Great Britain Journals Press Membership

Design of a Real-Time, Multilingual, Emotion-Aware Cyberbullying Detection System using Multi-Teacher Knowledge Distillation and Explainable AI

Prof. Dhananjay R Raut, Harsh J Sakpal, Madhur V Shinde, Aman S Singh & Vishal R Yadav

ABSTRACT

Social media cyberbullying has propagated rapidly and is being experienced by individuals worldwide. It tends to be expressed using sarcasm, emotional language, and multiple languages, making it difficult to determine the identity of the perpetrator. Although automated detection systems are becoming increasingly prevalent, the majority of existing systems suffer from language issues, function only in offline batch mode, and are black-box models that cannot be interpreted. These constraints make it more difficult to intervene with speed and transparency.

This paper offers a real-time, multilingual system for detecting cyberbullying, using explainable AI, emotion and sarcasm detection, and Multi-Teacher Knowledge Distillation (MTKD) to address shortcomings.

The system leverages an ensemble of transformer-based teacher models, like mBERT, XLM-R, and IndicBERT, to capture language-specific features.

Keywords: cyberbullying, real-time NLP, multi-teacher knowledge distillation, explainable AI, XGBoost, SHAP, emotion detection, sarcasm detection, multilingual NLP.

Classification: DDC Code: 006.35

Language: English

Great Britain Journals Press

LJP Copyright ID: 975841
Print ISSN: 2514-863X
Online ISSN: 2514-8648

London Journal of Research in Computer Science & Technology

Volume 25 | Issue 4 | Compilation 1.0

Design of a Real-Time, Multilingual, Emotion-Aware Cyberbullying Detection System using Multi-Teacher Knowledge Distillation and Explainable AI

Prof. Dhananjay R Raut[α], Harsh J Sakpal[σ], Madhur V Shinde[ρ], Aman S Singh[Ↄ] & Vishal R Yadav[¥]

ABSTRACT

Social media cyberbullying has propagated rapidly and is being experienced by individuals worldwide. It tends to be expressed using sarcasm, emotional language, and multiple languages, making it difficult to determine the identity of the perpetrator. Although automated detection systems are becoming increasingly prevalent, the majority of existing systems suffer from language issues, function only in offline batch mode, and are black-box models that cannot be interpreted. These constraints make it more difficult to intervene with speed and transparency.

This paper offers a real-time, multilingual system for detecting cyberbullying, using explainable AI, emotion and sarcasm detection, and Multi-Teacher Knowledge Distillation (MTKD) to address shortcomings.

The system leverages an ensemble of transformer-based teacher models, like mBERT, XLM-R, and IndicBERT, to capture language-specific features. Then, the models collaborate to produce a lightweight XGBoost classifier. To assist with the interpretation of context, additional layers are incorporated to identify sarcasm and emotion. SHAP (SHapley Additive Explanations) is employed to provide each prediction token-level interpretability. Algorithmic and architectural design of a system that would form a transparent, efficient, and deployable solution to detect cyberbullying in different emotional and linguistic contexts is the focus of this study.

Keywords: cyberbullying, real-time NLP, multi-teacher knowledge distillation, explainable AI, XGBoost, SHAP, emotion detection, sarcasm detection, multilingual NLP.

Author α σ ρ Ↄ ¥: Computer Engineering Watumull Institute of Engneering and Technology Thane, India.

I. INTRODUCTION

Cyberbullying refers to the act of sending injurious, discriminatory, or insulting messages through digital media. This has increased in India with widespread smartphone penetration, low-cost internet, and regular use of apps such as WhatsApp, Instagram, Twitter (X), and Facebook. Teenagers, women, minority groups, and celebrities are most affected, and victims experience emotional trauma, reputational harm, and occasionally self-injury or suicide [1].

India's language heterogeneity presents a subtlety problem for detection systems. Users tend to post in several regional languages—Hindi, Bengali, Tamil, Telugu, Marathi, Kannada, etc.—and also often mix these with English, creating code-mixed utterances such as Hinglish or Tanglish. These are frequently posted in Roman alphabet rather than native scripts, employing casual grammar and phonetics (e.g., "tu bahut irritating hai yaar").

Conventional NLP systems and deep learning models usually fail with such code-mixed or transliterated text. Moreover, bullying content tends to include sarcasm, cultural hints, slang, emojis, and emotive tone. For instance, "wah kya smartness hai" may be used as praise or sarcastic

London Journal of Research in Computer Science & Technology

insult depending upon context—this kind of nuance that most models fail to capture [2].

Current cyberbullying detection systems have significant drawbacks:

They are created for a single language (such as English or Thai) and don't address multilingual and code-mixed content that is common in India.

They run offline in batch mode, incapable of detecting toxic posts as they go up.

They use black-box neural models (such as CNN or LSTM), providing no insight into why a post was detected.

They don't have emotion or sarcasm analysis, decreasing accuracy in real-world scenarios [3].

To cover these loopholes, we introduce a real-time multilingual detection framework, designed specifically for Indian languages. The approach employs Multi-Teacher Knowledge Distillation (MTKD) through the ensembling of transformer models (like mBERT, XLM-R, and IndicBERT) as teacher models. The collective outputs of these teacher models are distilled to a lightweight XGBoost student model to support efficient and high-speed inference.

Two auxiliary layers are also added to further improve contextual comprehension:

An emotion recognition module powered by GoEmotions-BERT, which is fine-tuned over Indian code-mixed social media.

An in-house sarcasm detection module trained on both Reddit sarcasm data and Indian code-mixed instances.

To promote interpretability, the system incorporates SHAP (Shapley Additive Explanations) to detect and flag distinctive words that have impacted the decision. This transparency increases trust for human moderators while aiding ethical deployment.

This work is concerned with the system's architectural and algorithmic design—ranging from data flow, model choice, knowledge distillation, emotion/sarcasm fusion, and explainability modules—giving an action plan for future execution in various Indian content environments.

II. LITERATURE REVIEW

Automatic cyberbullying identification has been a topical area of study in the past decade. Numerous approaches have been suggested based on machine learning, deep learning, and NLP. Most of the prior work, however, is constrained in language generality, interpretability of the model, and real-time processing capability. This section provides an overview of five important areas that pertain to our suggested design: knowledge distillation, multilingual NLP models, emotion recognition, sarcasm detection, and explainable AI.

2.1 MTKD with XGBoost for Cyberbullying

Our project base work was presented by Sathit Prasomphan [4], in which a combination of Multi-Teacher Knowledge Distillation (MTKD) and XGBoost was suggested to detect cyberbullying in Thai social media. Several transformer models were used as teacher models to give soft probability outputs. These were distilled into a student XGBoost model to enhance efficiency.

Although the method had good precision, it was limited to Thai material and did not support features such as multilinguality, emotion detection, or explainability. Furthermore, it was developed for static data instead of real-time.

2.2 GoEmotions – Emotion Detection using BERT

GoEmotions is a large dataset developed by Google with more than 58,000 Reddit comments annotated with 27 emotion categories and a neutral category. High accuracy in emotion classification has been demonstrated by a fine-tuned BERT model on this dataset [5]. It comes handy in interpreting the tone of text — if it conveys anger, happiness, sadness, etc.

But GoEmotions was built primarily for English text and is not designed for online abuse or cyberbullying detection. Furthermore, the model has not been implemented in multilingual or

code-mixed environments such as are typical in India.

2.3 SHAP -Explainable AI for NLP

SHAP (SHapley Additive Explanations) is a widely used framework to explain the predictions of machine learning models. It provides a contribution score to each word or token towards the final prediction [6]. SHAP has been applied to various fields like healthcare and finance to increase model transparency.

SHAP is primarily used in NLP with classifiers such as XGBoost or BERT. But there has not been a lot of work on SHAP for cyberbullying detection, particularly in multilingual or affect-based scenarios. Furthermore, SHAP explanations are computationally costly, which makes them less practical to use in real-time

2.4 Sarcasm Detection using Deep Learning

It is challenging to identify sarcasm since it tends to rely on context and implied meaning. Researchers such as Mishra et al. [7] have employed BiGRUs, LSTMs, and CNNs that were trained on Reddit or Twitter to recognize sarcastic posts. The models are moderately successful but tend to be trained on English data alone.

There is no current model that entirely enables real-time sarcasm detection in Indian code-mixed languages. In addition, sarcasm detectors are typically standalone and not incorporated into cyberbullying detection systems.

2.5 RoERTa for Toxiuc Language Classification

RoBERTa is a state-of-the-art variant of BERT and has been extensively applied in toxicity detection tasks, such as datasets like Jigsaw Toxic Comments. It provides excellent accuracy in classifying hateful or offensive speech [8].

Nonetheless, RoBERTa is heavy on resources, opaque, and not optimized for real-time inference. It also does not support multiple Indian languages or emotional and sarcastic content.

2.6 Comparitive Summary of Prior Work

Sr. No.	Paper/Approach	Method	Limitations
1	MTKD with XGBoost [4]	Knowledge Distillation + XGBoost	Thai-only, offline, no emotion or XAI support
2	GoEmotions [5]	Emotion-labeled BERT model	English-only, not focused on bullying
3	SHAP NLP [6]	Word-level explanation for XGBoost	Not integrated with cyberbullying pipeline, high computational cost
4	Sarcasm Detection [7]	BiGRU and CNN on Reddit/Twitter	English-only, not real-time or multilingual
5	RoBERTa Toxicity [8]	Toxic comment classification	Black-box model, lacks emotion/sarcasm modules and multilingual support

III. CONCLUSION OF LITRATURE REVIEW

The review vehemently points out that none of the available systems provide real-time, multilingual, emotion-sensing, and explainable cyberbullying detection in an integrated manner. Particularly in a culturally and linguistically diverse nation like India, where language, emotion, and sarcasm combine in intricate manners, available solutions are inadequate.

The system proposed tries to bridge this gap by:

- Utilizing MTKD for ensemble of multilingual teacher models,
- Including SHAP for explainability purposes,
- Combining emotion and sarcasm detectors for improved context,
- Utilizing a light XGBoost model for efficient inference.

This combined design offers a scalable and deployable backbone for abusive content detection over India's diverse linguistic and social terrain.

IV. PROBLEM STATEMENT

Cyberbullying has emerged as a developing issue in India with the upsurge of online activity on social media like Twitter, Instagram, Facebook, and WhatsApp. In contrast to physical bullying, cyberbullying can take place at any time and from anywhere and even anonymously—resulting in long-term psychological trauma, particularly among young people, women, and marginalized groups [9]. In spite of great improvement in natural language processing (NLP) and machine learning, the existing cyberbullying detection mechanisms are unable to meet the real-world requirements of India's multilingual and culturally rich internet population.

The most significant challenge among them is diversity of languages and code-mixing. Indian users typically write in Hindi, Tamil, Bengali, Telugu, or Marathi—or code-mix them with English (e.g., Hinglish or Tanglish). Such posts are usually composed in Romanized script with heavy usage of non-standard grammar, abbreviations, emojis, and web slang (e.g., "Tu kya bakwaas kar raha hai bro ????♂"). The standard monolingual models learned on formal English data are not able to handle such noisy and informal content efficiently [10].

Secondly, the majority of current systems operate offline in batch mode, analyzing pre-gathered datasets [1]. This introduces a lag between when something is posted and when it's analyzed—making the system useless for sites that require real-time moderation. Without instantaneous discovery, bullying or toxic comments can spread virally before any moderation is done, amplifying its damage.

One of the most important gaps is explainability. Recent transformer-based models such as BERT and RoBERTa provide outstanding NLP accuracy but are black-boxes, providing no or minimal information on why a choice was made. This transparency issue is concerning in legal, educational, or institutional environments where decisions must be justified, confirmed, or audited [6][8].

In addition, the majority of systems do not consider emotional tone and sarcasm, which are particularly prevalent in Indian online discourse. Such a statement as "Waah kya sanskaar hai!" might be an honest compliment or caustic sarcasm depending on the situation. With no consideration for sarcasm or measurement of emotional intensity, systems can either fail to notice problematic content or produce false positives, eroding trust in automated moderation tools [5][7].

Considering these constraints, the need for a real-time, emotion-sensitive, explainable, and multilingual cyberbullying detector, specifically for Indian users, is highly urgent. The system should:

- Be able to handle multilingual and code-mixed text input, particularly from Indian languages
- Function in real time so continuous monitoring and instant alerts are possible
- Detect emotion and sarcasm to enhance contextual categorization
- Be explainable so that human moderators can comprehend and rely on model decisions
- Be light-weight and efficient, allowing deployment in real-world applications

To fill this gap, we present a design that employs Multi-Teacher Knowledge Distillation (MTKD) to merge the strengths of diverse transformer models like IndicBERT, mBERT, and XLM-R, all trained on varying language domains. The teachers impart their soft-label knowledge into an efficient and light-weight XGBoost student model, which is deployable for real-time prediction. We also add emotion and sarcasm detection layers to understand user tone and intent, and lastly apply SHAP (SHapley Additive Explanations) to make each decision interpretable on the token level.

This combined system—tuned to India's digital linguistics—seeks to greatly enhance the detection of toxic behavior across social networks and make detection such that it becomes actionable, ethical, and transparent.

V. OBJECTIVES

The main aim of this project is to create a real-time, multilingual, emotionally intelligent, and explainable cyberbullying detection system custom-made for the nuances of Indian social media. The following is the system's particular objectives in a precise way, ranging from the entire architecture to linguistic diversity, affective comprehension, and moderator interaction design.

5.1 Real-Time Cyberbullying Detection

The system to be proposed is such that it runs in real-time, processing posts upon posting. Contrary to batch-processing models that work on static datasets after they are gathered, this architecture makes use of the Tweepy (Twitter) and PRAW (Reddit) APIs to scan social feeds in real-time.

For instance, when someone tweets "You're such a burden, nobody wants you here," the system needs to process immediately, classify, and alert the moderators in a matter of seconds. Early warning is essential to avoid escalation, especially in high-risk scenarios with youths or vulnerable communities [1], [9].

5.2 Processing Indian Multilingual and Code-Mixed Content.

Considering India's linguistic diversity, the users tend to switch between languages frequently (e.g., "Tum kya bakwas kar rahe ho?" in Hinglish). To deal with this, the system is equipped with:

- Multiple Indian languages: Hindi, Tamil, Bengali, Telugu, Marathi.
- Code-mixed and Romanized scripts.

This is accomplished through IndicBERT, mBERT, and XLM-R, each trained on multilingual corpora with the ability to comprehend local phonetics, grammar variation, and transliteration. These models assist in identifying offensive content for different linguistic inputs [2], [10].

5.3 Multi-Teacher Knowledge Distillation (MTKD)

The architecture employs a Multi-Teacher Knowledge Distillation method in which multiple high-performance transformer models serve as teachers. Each teacher is fine-tuned on a target language or code-mixed data and produces soft probabilities as output.

For example, IndicBERT is good for Indian regional scenarios and XLM-R is good with low-resource multilingual data. Such outputs are consolidated (frequently through weighted average taking reliability of models into consideration) and distilled into a student model, enhancing multilingual generalization without sacrificing speed [1].

5.3 Lightweight XGBoost Student Model

The end student model is an XGBoost classifier, selected because it has:

- Low latency
- High interpretability
- Simple integration with SHAP for explainability

This model receives the distilled soft targets and is tuned using grid search (e.g., max_depth, learning_rate, lambda) to maximize F1-score. It can detect severe phrases like “Go away forever. You’re useless” with minimal computational overhead—ideal for real-time deployment [1], [6].

5.4 Emotion Detection Layer

Emotions are an essential aspect of cyber-abuse, particularly of indirect bullying. This framework incorporates an emotion classifier trained on GoEmotions-BERT, which projects each post to one out of 27 emotion categories: anger, sadness, and fear.

A subtle post like “I’m tired of pretending to be happy ” may not be toxic, but signals distress. This classifier, adapted for code-mixed Indian language content, flags emotionally vulnerable posts to ensure protective action [5].

5.5 Sarcasm Detection Layer

Indian social media often contains sarcasm that masks hostility. Consider a sarcastic comment like

"Wah kya smartness hai!", which can be both humorous and derogatory depending on context.

A BiGRU or attention-based LSTMs sarcasm classifier is employed, which is trained on labeled sarcasm data from Twitter and Reddit. The module detects sarcastic posts based on language indicators, emojis, and context shifting, greatly improving the accuracy of classification [7].

5.6 Explainable AI (SHAP-based)

Model explainability is facilitated through SHAP (SHapley Additive Explanations). SHAP calculates token-level attribution, allowing moderators to know why a post was detected.

For instance, in the statement "Nobody wants you around anymore", SHAP will underline "nobody" and "anymore" as primary triggers. These visual justifications increase transparency and establish trust, particularly in moderation environments that mandate human justification or auditing [6].

5.7 Severity Classification of Cyberbullying

Posts are categorized into three levels of severity based on their content:

- "You are annoying" → Mild
- "You should disappear" → Moderate
- "You deserve to die" → Severe

This is attained through the use of a mix of toxicity scores, emotion classes, and sarcasm indicators. By introducing severity levels, the system allows moderators to prioritize high-risk content and act accordingly [3], [9].

5.8 Moderator Dashboard for Monitoring and Action

To aid human moderators, a ReactJS dashboard is created. It shows flagged posts with:

- Detected language
- Emotion
- Sarcasm
- SHAP explanation
- Severity level

They may take steps like "Ignore," "Report," "Delete," or "Export logs." This interface offers complete transparency, accountability, and usability to platform teams, educators, or legal authorities that need to track abuse patterns [11].

VI. PROPOSED DESIGN

6.1 Data Collection and Sources from Social Media

To create a good system for finding cyberbullying, we need lots of different kinds of data we can trust. Social media sites like X, Facebook, Instagram, and YouTube have tons of content made by users, and that's where bullying happens. We start by grabbing posts and comments using official tools or data sources we pay for. For example, we can use X's API to grab tweets that have certain words or hashtags linked to bullying, like loser, hate, or idiot, or mentions of user accounts. We can also get data from Reddit and YouTube comments using tools that are available to everyone. We make sure to keep user data private by removing personal info and keeping everything safe.

In addition to open-source datasets, curated code-mixed language corpora—especially from Indian social media—are included. Users often mix English, Hindi, Bengali, Tamil, and other languages within the same post, which complicates detection. Many Hindi words also appear in Roman script (for example, "ch-**iya") or as slang variants, raising the difficulty. To address this, the dataset includes both monolingual and code-mixed samples across multiple languages to reflect diverse community contexts.

Throughout, ethical standards are applied by anonymizing personal information such as names, email addresses, and phone numbers. Because bullying instances are relatively rare, data augmentation techniques like synonym replacement, back-translation, and paraphrasing are used to expand the dataset. This reduces bias toward non-bullying examples and improves the model's ability to generalize across content types.

6.2 Preprocessing and Cleaning

Online posts can be a mess. People misspell words, use emojis to show feelings, and often use

abbreviations or loose grammar. If you don't clean things up first, it's easy for a model to miss teasing or bullying. This part gets the text ready but keeps the original meaning.

- Noise and Slang Removal – We fix things like umm, lol, and btw. We also squeeze repeated letters in words like looooser and make them loser. We keep a list of slang terms (including bad ones from different areas) and replace them with standard markers, so the models understand them better.
- Emoji Expansion – Emojis carry clear signals: "😂" can hint at sarcasm, while "😡" suggests anger. Following Felbo et al. [12], the pipeline maps emojis to short text descriptors (e.g., 😂 → "laughing-crying") so text-only models still capture affective cues.
- Language Identification & Romanized Text Handling – Many Indian users write Hindi or Tamil in Roman script, so a language ID step labels tokens as English or a regional language. Then, we turn words back to native scripts with tools like IndicTrans [13], ensuring slang like "madarchod" is recognized despite spelling variants.
- Normalization & Tokenization – We swap URLs, hashtags, and user names for tags like <URL> and <USER> to reduce clutter but keep post structure. We split the text into smaller bits (tokens) with tools made for multiple languages—this lets models learn from cleaner data.

Together, these steps turn noisy, informal text into clean, structured, and semantically rich sequences that are ready for deep representation learning.

6.3 Multi-Teacher Ensemble (mBERT + XLM-R + MuRIL)

Understanding the many ways people express themselves in different languages takes more than a single model. Instead of relying on just one, we bring together several powerful language models, each trained for multilingual tasks. This "multi-teacher ensemble" blends their strengths to make our system both broader and more dependable.All the teacher models are trained on the same dataset, and instead of just using hard labels, we merge their soft probability outputs to gain richer insights. This method preserves uncertainty and allows us to better understand cross-lingual details.

- mBERT (Multilingual BERT) [14] – Supports over 100 languages with strong cross-lingual capabilities for various NLP tasks. Its subword tokenization helps navigate complex morphology in Indian languages and reduces out-of-vocabulary issues.
- XLM-RoBERTa (XLM-R) [15] – Trained on an extensive multilingual corpus, it excels particularly at zero-shot transfer across languages. It's especially effective for context-rich cues, like cross-lingual sarcasm, offering strong, language-agnostic contextual embeddings.
- MuRIL (Multilingual Representations for Indian Languages) [16] – Built specifically with Indian languages and code-mixing in mind, including Romanized text. This makes it ideal for identifying harassment or bullying in mixed scripts such as Hinglish and Tanglish.

The ensemble combines the soft predictions from each teacher to use their complementary strengths and minimize blind spots. For example, if mBERT outputs a high likelihood for "hate" while MuRIL detects a Roman-script Hindi slur, the combined result offers a more reliable understanding than either model alone.

6.4 Knowledge Distillation

While teacher models tend to be highly accurate, they often require major computational resources, making them less suitable for real-time applications. To address this challenge, we use knowledge distillation (KD) [17]. This technique involves using the softened probability outputs of a large teacher ensemble to guide the training of a smaller, more efficient student model.

The core idea is that the ensemble of teachers doesn't just say 'yes' or 'no' to bullying—instead, it gives nuanced scores (for example, "70% bullying, 20% sarcasm, 10% neutral"). The student model then learns from these slightly softened, richer targets, instead of hard yes/no labels. This lets it

understand more about the grey areas between classes.

During training, the student tries to match the teachers by minimizing the difference (using KL divergence) between their predictions. We also use "temperature scaling" to smooth the probabilities, helping the student grasp a wider range of outcomes. In the end, the student model learns not only to make correct classifications but to mirror the deeper understanding of its teachers—while running much faster on ordinary computers.

6.5 Student Model (XGBoost Classifier)

After distillation, we pass the learned information to a lightweight XGBoost classifier. XGBoost is fast, easy to understand, and runs well even on basic hardware (like school desktops). It takes a mix of raw transformer outputs—summarized into embeddings—and a handful of hand-crafted features to spot bullying in social media posts. These features include:

- Term frequency–inverse document frequency (TF-IDF): How rare or common each word is in the dataset.
- Sentiment polarity scores: Does the post "sound" positive, negative, or neutral?
- Stylometric features: Punctuation counts, upper/lowercase ratios, number of exclamation marks, etc.
- Code-mixing index: For example, what percent of a post is in Hindi versus English?

By combining deep text understanding with these simple, readable features, XGBoost gives us predictions that are reliable and still easy to explain. Because it runs efficiently on CPUs, it's practical to use for real-time checks on everyday school computers.

6.6 Emotion & Sarcasm Detection

Bullies don't always come out and say what they mean—instead, they often hide behind sarcasm or emotional language. That's why our system includes two special detectors:

- Emotion Detector — Built on GoEmotions-BERT [], this model classifies messages into emotions like anger, sadness, joy, disgust, or fear. Bullying is often linked to strong negative emotions (especially anger and disgust), so detecting these feelings helps us judge how serious a message is.
- Sarcasm Detector — Sometimes, a message looks positive but actually carries a cutting or mean tone—for instance, "Wow, you're such a genius 🙄." Our sarcasm detector tackles this using a BiGRU with attention network [19], which has been trained on real sarcastic posts (and uses emoji cues, too). The model "pays extra attention" to words and symbols that hint at sarcasm, making it easier for us to spot when someone is being sneaky.

The outputs from both detectors feed into XGBoost's final decision, making it much less likely that hidden or subtle bullying slips by unnoticed.

6.7 Final Decision Layer

The final decision brings together the student model, emotion detector, and sarcasm detector all send their findings into a final decision layer, which produces a single, easy-to-use result. The layer returns three outputs:

- Binary Decision – A simple Yes/No on whether the post is considered bullying.
- Severity Score – A number between 0 and 1, showing how intense the bullying is (so you can decide, for example, if a warning is enough or if a block is needed).
- SHAP Explanation [6] – For full transparency, the system uses SHAP (SHapley Additive exPlanations) to break down which words or features contributed most to the final call—for example, the score might show that a certain abusive word or sarcastic emoji tipped the scales.

This kind of explainability helps moderators trust the system, makes decisions fairer, and fits best practices for ethical AI.

6.8 Moderator Dashboard

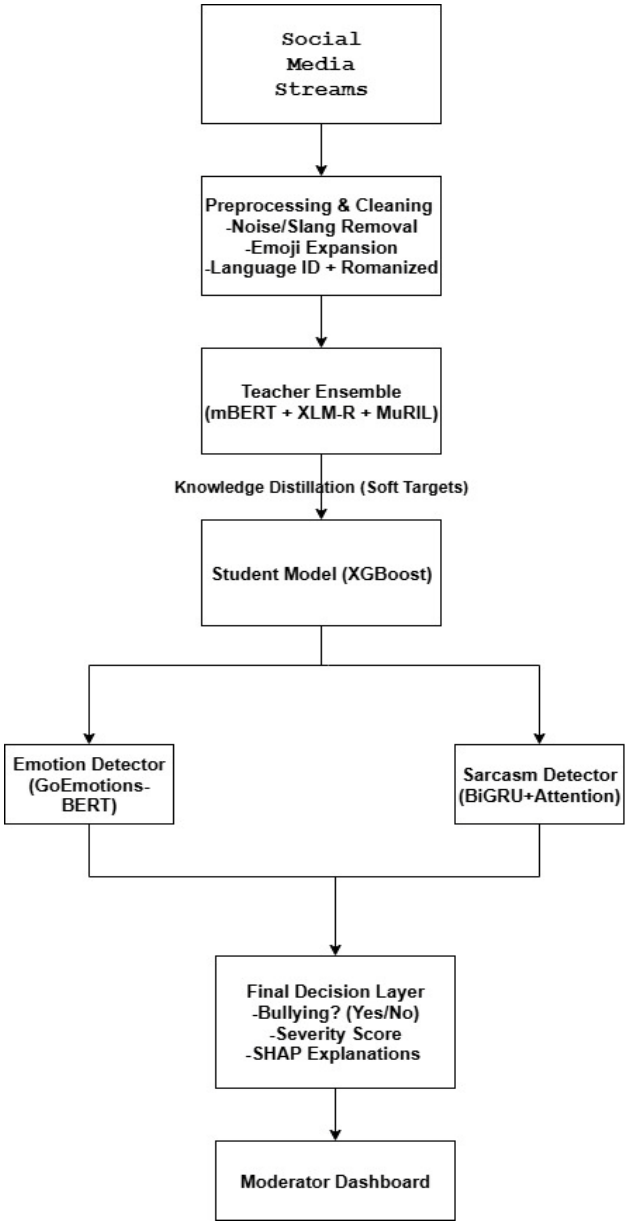
After processing, a Moderator Dashboard turns complex outputs into simple, actionable views. It

clearly indicates what requires immediate attention and offers sufficient context to enable quick, confident decisions.

- Flagged Content Viewer – Shows the original post and clearly marks which parts are abusive, making it easy to spot and review problem language fast.
- Severity Analytics – The dashboard brings trends to life: track which terms, languages, or types of abuse are flaring up over time, and catch new slang or patterns before they spread.
- Real-Time Alerts – If something urgent—like a threat or hate speech—shows up, the dashboard pings the moderator right away. That way, high-severity issues never slip through the cracks.
- Explainability Reports – When a post gets flagged, moderators aren't left guessing. The dashboard lays out a straightforward summary (thanks to SHAP), showing exactly why the system flagged that message.

The goal here is to keep the human in charge. By keeping controls simple and explanations easy to follow, the dashboard supports ethical, transparent content moderation—helping moderators protect their communities and make decisions with confidence [9].

A. System Workflow Diagram.



VII. TOOLS AND TECHNOLOGY

7.1 Social Media API Integration (Free Alternatives)

To keep costs down and still collect enough data, the system leans on free API options, with pacing and batching to avoid rate caps.

- Twitter (Tweepy with free API v2 access)
  - Free developer tiers allow limited monthly volume (policy-dependent, often a few hundred thousand tweets).
  - Real-time filters by keywords or hashtags make targeted streaming easy.
  - JSON output drops straight into preprocessing and NLP steps.
- Reddit (PRAW)
  - Works free with personal/app credentials within rate limits.
  - Streams or polls posts and comments from chosen subreddits tied to bullying detection.
- Other free platforms
  - Pushshift API: Handy for historical Reddit backfill and exploration, availability varies.
  - Mastodon API: Federated, open-source streaming from selected instances.
  - Facebook/Instagram alternatives: Crowd Tangle offers limited free research access for public data.

Run short, staggered jobs (every 10–30 seconds) to stay within free limits, and write line-delimited JSON from small Python scripts for durable, auditable NLP pipelines [20].

7.2 Language Detection and Handling.

Free tools for multilingual detection and processing:

- langdetect (Python): Open-source, detects 55+ languages, and works with simple rules to handle code-mixed text.
- langid.py: A lightweight, fully offline detector that's easy to run in scripts or on devices.
- Indic NLP Library: Tokenizes and provides utilities for Indian languages to produce clean inputs.
- IndicTrans: Free transliteration from Romanized text to native scripts (e.g., Hinglish → Hindi).

Example:

```
from langdetect import detect
text = "Tum kya kar rahe ho bro?"
lang = detect(text)  # returns 'hi' for Hindi
```

Free Handling for Indian Languages:

- Hindi, Tamil, Bengali → By using IndicTrans + IndicNLP.
- Code-mixed → Combine outputs from mBERT + XLM-R (free HuggingFace models).

7.3 NLP Processing (Free Libraries)

It uses open-source libraries for all NLP steps.

- NLTK: It handles tokenization, stopword removal, and sentence splitting.
- SpaCy: Fast tokenization and dependency parsing for production use.
- PyThaiNLP: This is optional and helps with Thai benchmarks or multilingual tests.
- Emoji (Python package): Maps emojis to text so sentiment can be captured.
- Slang normalization dictionaries: It uses free, custom mappings to normalize social media slang.

7.4 Transformer Teacher Models (Free and Open-Source).

The MTKD framework uses free pretrained transformer models available on HuggingFace.

- mBERT: bert-base-multilingual-cased.
- XLM-RoBERTa: xlm-roberta-base.
- MuRIL: Free Indian multilingual model for Hindi, Bengali, Tamil, and Telugu.
- ThaiBERT: This is optional and used for benchmarking.

All models are publicly available, with no paid subscription required [23][24].

7.5 Knowledge Distillation Framework

It relies on free and open-source tools for training and inference.

- XGBoost (Python): Free, open-source tree boosting library.
- scikit-learn: It handles preprocessing, train–test splits, and metrics.
- NumPy / pandas: These are used for data handling.

Training process:

- It distills soft labels from the teacher models.
- It takes a weighted average by F1-score per language.
- It trains XGBoost on the distilled labels for fast inference.

7.6 Emotion And Sarcasm Detection (Free)

- GoEmotions-BERT: It uses a HuggingFace model trained on 28 emotion classes; this is free to download and fine-tune.
- BiGRU Sarcasm Detection: It uses a Keras/TensorFlow implementation; both frameworks are free.
- Datasets: It uses free Hindi-English tweet datasets from Kaggle or research repositories for training.

7.7 Explainability with SHAP

- SHAP library: This is a free, Python-based interpretability tool.
- Generates token-level explanations for XGBoost predictions.
- It is critical for moderator transparency and ethical AI.

7.8 Backend Infrastructure (Free Options)

- FastAPI: Free, async Python framework for REST APIs.
- Flask: It is a lightweight alternative for quick prototypes.
- Server Hosting:
  - Railway Free Tier: It offers about 500–1000 hours per month.
  - Render Free Tier: It offers about 750 free hours per month.
  - Replit: Supports free prototyping and hosting.

7.8 Frontend and User Interface (Free)

- ReactJS: It is free and open-source library for building user interfaces.
- Charts/Heatmaps: react-chartjs-2 or Plotly.js, both free.
- Deployment: Firebase Hosting (free tier) for static React apps.

7.9 Database and Hosting (Free Alternatives)

MongoDB Atlas (Free Tier): 512 MB storage; a good fit for JSON documents.

- Firebase Firestore (Free Tier): Up to 50,000 reads/writes per day.
- Hosting: Free tiers on Railway, Render, or Firebase Hosting for small deployments.

VIII. EXPECTED OUTCOMES

The proposed approach to multilingual cyberbullying detection in code-mixed Indian languages is expected to produce outcomes across multiple technical, practical, and social levels. The outcomes will be divided into four main areas: (1) system design and overall project outcomes, (2) technology and methods, (3) deployment and project management, and (4) societal and user impact.

8.1 Systems Design and Final Project Outcomes

This project will be based on modular and scalable architecture that integrates the data source and acquisition, pre-processing, model training, classification, and visualization layers.

- User Interface (UI): The principal interface will be a web dashboard built with React (frontend) and FastAPI (backend) that will enable monitoring of real-time social media streams from Twitter, Reddit, YouTube, and Facebook. Importantly, it will have multilingual input support for English, Hindi, Hinglish, Bengali, Tamil, and Indic Languages, providing support for communities across various languages.
- Data Pipeline: Data will be obtained using free APIs for Twitter (via Tweepy [20]), Reddit (PRAW [29]), Pushshift [30], and YouTube Data API v3 [31]. Furthermore, public reports

(such as from the Facebook Transparency Center [32]) will also be appended to use as published benchmark references for harmful content detection all at once. The preprocessing pipeline will consist of tokenization, transliteration, and emoji treatment, primarily using IndicNLP [21] and emoji libraries [25] library, so that the system can detect sarcasm, codeswitching, and the emotion of a situation in context.

- Visualization and Alerts: The dashboard will be designed as a professional moderation console. The left panel will show feeds of live data streams (tweets, posts, comments); the middle panel will display classification results; and the right panel will describe the Bystander Bullying scale, severity, toxicity categorizations, and emotional intensity output. Alerts will be sent to the moderators when the system detects high-risk bullying content.

8.2 Technical and Methodological Outcomes

The technical outcomes will focus on enhancing detection and monitoring accuracy; improving efficiency of model performance; and explainability in multilingual and code-mixed contexts.

- Enhanced Multilingual Detection: The improved multilanguage detection process will deploy pre-trained transformers, such as MuRIL [16] (a multilingual variant of BERT), IndicTrans [13] (a multi-script transformer appropriate for Indic languages), and RoBERTA [8], which will have all subsequently outperform traditional monolingual baselines. The inclusion of emoji in detection [12] will also improve sentiment detection performance, especially within sarcasm or humour-orientated bullying context.
- Evaluation Metrics: In addition to accuracy, recall, and F1-score [18], the system will utilize ROC curve analysis [33] to present the trade-offs of values between sensitivity and specificity for more reliable, actionable outcomes in high-stakes decision making.
- Comparative Model Benchmarks: The results of the proposed pipeline will be compared to the most recent abusive language detection models such as HateBERT [34] to show how pre-trained domain-specific models improve realizable outcomes in abusive and cyber bullying context.
- Explainability: To enhance trust and incentivize adoption, explainable AI [6] processes will be incorporated in which words, phrases, or emojis are highlighted as influences that provoke bullying classification. For example, in a Hinglish tweet that states “Tu loser hai 😂”, the system will identify “loser” as a related toxic word and that the laughing emoji as reinforcement sarcastically.
- Efficiency through Knowledge Distillation: Lightweight transformer-based models [17] will be trained to mimic the performance of large models while requiring less resources, creating a knowledge transfer model that can be utilized in data-poor settings.

8.3 Project Execution and Development

The project will encompass a complete step-by-step procedure, from data collection through to live deployment to be a scalable end to end solution.

- Data Collection and Storage: Where data is collected from Twitter [20], Reddit [29] and YouTube [32] APIs, data will be maintained in either MongoDB Atlas [27] or Firebase Firestore [28] in support of scaling. Data normalization and cleaning will be conducted in preprocessing queues, prior to feeding the model.
- Training and Evaluation: Training will utilize either Google Colab [26] or Kaggel [21] cloud services with the support of any available GPU hardware. Evaluation will involve multi-class classification metrics supplemented with ROC analysis [33].
- Deployment Infrastructure: The backend model will be deployed via FastAPI [26], whereas the web dashboard will connect with REST APIs. Demonstrations for the public or academic instances, would utilize Heroku’s [35] free cloud-hosting service and therefore eliminate the need for expensive infrastructure.

- Monitoring and Human-in-the-Loop Feedback Loop: A feedback loop will log moderation actions taken by the moderators in order to facilitate re-training suggesting models learn from adapting slang, changes in cultural variances, and detect emerging bullying patterns.

8.4 Social and User-level Outcomes.

Any effects at the societal and user level are equally as important as the technical deliverables of the system.

- Intervening Early to Online Harassment: The system will help mods, NGOs, schools, and online platforms catch early stages of harassment and act on it early on, as prolonged exposure to online harassment has substantial psychological and emotional costs [9].
- Safer and More Inclusive Digital Environment: Organizations including schools and workplaces will be able to implement the system in order to protect vulnerable groups, notably children/teens. This is also consistent with the UNESCO "Feasibility Study for Safe and Inclusive Digital Learning Spaces for Children" [36].
- Supporting Multilingual Communities: The system will cater to the many Indic languages and code-mixing prevalent in South Asian communities, creating informal collective leverage. This also closes the gap of moderating in pre-dominantly monolingual systems [2].
- Different Applicability Use-Case Scenarios: The solution will be applicable to more than just social media moderation, as it will also have applications in education, workplace harassment, and law enforcement to name a few.
- Contributing to Research and Open Resources: The projects data sets and trained models will be published as multilingual open-access research resources for the academic and industry to innovate on [15].

IX. FUTURE SCOPE

Detecting cyberbullying in multilingual and code-mixed contexts is a burgeoning area of research and the current work provides a number of pathways for future work.

9.1 Extending to More Languages and Dialects

The current system works with English, Hindi and a subset of Indic languages. Future expanding this framework to other regional dialects and minority languages, such as Assamese, Odia, and Konkani, has benefits, particularly because many of these languages have few or no labelled datasets [2]. With the increasing availability of cross-lingual transfer learning techniques [13], the thinking would be that those scarce languages could be modelled much easier, without large annotated corpora, but with some distance from the original dataset.

9.2 Syncing with Audio and Video Streams

This study primarily concerns itself with the text data related to cyberbullying. Future work may explore what it takes to bulk out the system, in a multimodal way, with YouTube videos, live audio chats, and memes. Work integrating APIs, such as the YouTube Data API v3 [31], and even speech-to-text technologies in real time should be explored, to collect comments, transcripts, and captions as they occur. The multimodal approach will strengthen detection on platforms that provide bullying in multiple formats.

9.3 Real-Time Deployment at Scale

This implementation illustrates a scalable cloud deployment of FastAPI [26] on Heroku [35]. Additional work could focus on edge deployments for mobile devices and interconnection of moderation pipelines of social platforms. This enables real-time detection of abusive content and flagging it and removing it before the use becomes widespread.

9.4 Contextual and Psychological Analysis

Cyberbullying detection cannot be limited to keyword or sentiment analysis. The future scope should provide, in addition, some form of

psychological perspectives and behavioral modelling to distinguish between banter, sarcasm, or real bullying [12]. This can be accomplish when offer the expertise of psychologists, educators and NGOs [36] or a more human-centered moderation.

9.5 Adaptive Learning with Continuous Feedback

The models may become obsolete as internet slang, memes, and emoji use develop and change. Therefore, masters of all Interactive Learning can be embedded in the future with a feedback link so moderators' decision can automatically retrain models and provide ongoing adaptability [6].

9.6 Other Application Domains

There are additional domains, other than social media, this framework could be applied to:

- Education: Monitoring forum or chats in classrooms are the right environment and opportunity to inform students they are not subjected to online harassment.
- Work settings: Employees should be assured their corporate communication tools are free from bullying and harassment.
- Law enforcement: To monitor that organized cyber harassment and campaign hate or negative organizations are not beginning to take shape [34].

9.7 Engagement with Global Digital Safety Initiatives

Future systems can collaborate with programs led by UNESCO [36] and other digital safety initiatives to build global databases of annotated bullying data. This would support greater model quality and cross-jurisdictional collaboration to combat cyber harassment.

X. CONCLUSION

This project proposed a multilingual cyberbullying detection system for the heterogeneity both linguistically and culturally of India, with features such as code-mixed language, the use of emojis, and sarcasm detection. Implementation drew on plug-in free APIs for real time data acquisition (Twitter [20], Reddit [29] and YouTube [31]), state-of-the-art natural language processing (NLP) models (MuRIL [16] and HateBERT [34]), and cloud deployment options (Heroku [35]). The result is a well-defined system that is built to be robust and scalable.

The contributions of this work are summarized as follows:

1. A composite pipeline of cyberbullying detection in multilingual and code-mixed space.
2. Apart from sample rates presented using the baseline models, the translator-based models that lead to improvements can see significant increases in detection rates.
3. A dashboard that provides the moderator with real-time data with visual representation for alerts.
4. The work meets international standards for online safety (UNESCO [36]), which gives the work immediate topicality.

On the societal side, the project highlighted an AI-powered systems can be a source of promoting a safer and more inclusive digital scenario by its use with vulnerable populations, in particular children, adolescents and minority groups [9], [36].

To conclude this study, the focus has been to address the technical side of the multilingual cyberbullying and detection problem, with more importance placed on the deployment and social impact. As the project grows to involve more multimodal, multimodal learning and global engagement, the project has the ability to be a worldwide model for online safety.

REFERENCES

1. S. Prasomphan, "Enhance Social Network Bullying Detection Using Multi-Teacher Knowledge Distillation With XGBoost Classifier," *IEEE Access*, 2025. Available: https://www.researchgate.net/publication/392212453_Enhance_Social_Network_Bullying_Detection_using_Multi-Teacher_Knowledge_Distillation_with_XGBoost_Classifier
2. A. Patra, A. Das, B. Das, and S. Das, "Sentiment Analysis of Code-Mixed Indian

Languages: SAIL Code-Mixed Shared Task," arXiv preprint arXiv:1803.06745, 2018. Available: https://arxiv.org/abs/1803.06745
3. S. Mehendale, D. Dodia, and H. Palshetkar, "A Review on Cyberbullying Detection Using Machine Learning in English and Hinglish," *SSRN*, 2022. Available: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4116153
4. S. Prasomphan, "Enhance Social Network Bullying Detection Using Multi-Teacher Knowledge Distillation With XGBoost Classifier," *IEEE Access*, 2025. [Online]. Available: https://www.researchgate.net/publication/392212453
5. D. Demszky, D. Movshovitz-Attias, J. Ko, A. Cowen, G. Nemade, and S. Ravi, "GoEmotions: A Dataset of Fine-Grained Emotions," *arXiv preprint arXiv:2005.00547*, 2020. [Online]. Available: https://arxiv.org/abs/2005.00547
6. S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in *Advances in Neural Information Processing Systems*, vol. 30, 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf
7. A. Mishra, A. Jain, and P. Bhattacharyya, "A Deep Learning Approach to Sarcasm Detection in Social Media," *arXiv preprint arXiv:1605.01159*, 2016. [Online]. Available: https://arxiv.org/abs/1605.01159
8. Y. Liu, M. Ott, N. Goyal, et al., "RoBERTa: A Robustly Optimized BERT Pretraining Approach," *arXiv preprint arXiv:1907.11692*, 2019. [Online]. Available: https://arxiv.org/abs/1907.11692
9. UNICEF India, "UNICEF calls for concerted action to prevent online bullying," *Press Release*, 2021. Available: https://www.unicef.org/india/press-releases/safer-internet-day-unicef-calls-concerted-action-prevent-bullying-and-harassment
10. V. Srivastava and M. Singh, "Challenges and Considerations with Code-Mixed NLP for Multilingual Societies," *arXiv preprint arXiv:2106.07823*, 2021. Available: https://arxiv.org/abs/2106.07823
11. K. Maity, R. Jain, P. Jha, and S. Saha, "Explainable Cyberbullying Detection in Hinglish: A Generative Approach," *IEEE Transactions on Computational Social Systems*, vol. 11, no. 3, pp. 3338–3347, 2024. [Online]. Available: https://doi.org/10.1109/TCSS.2023.3333675
12. C. Felbo, A. Mislove, A. Søgaard, I. Rahwan, and S. Lehmann, "Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm," *Proc. Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1615–1625, 2017. [Online]. Available: https://aclanthology.org/D17-1169
13. A. Bapna et al., "IndicTrans: An effective transformer-based model for English–Indic translation," *Proc. 2022 Conf. North American Chapter of the ACL: Human Language Technologies*, 2022. [Online]. Available: https://aclanthology.org/2022.naacl-main.58
14. J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *Proc. NAACL-HLT*, pp. 4171–4186, 2019. [Online]. Available: https://aclanthology.org/N19-1423
15. A. Conneau et al., "Unsupervised cross-lingual representation learning at scale," *Proc. 58th Annual Meeting of the ACL*, pp. 8440–8451, 2020. [Online]. Available: https://aclanthology.org/2020.acl-main.747 K. Khanuja et al., "MuRIL: Multilingual representations for Indian languages," arXiv preprint arXiv:2103.10730, 2021. [Online]. Available: https://arxiv.org/abs/2103.10730
16. K. Khanuja et al., "MuRIL: Multilingual representations for Indian languages," arXiv preprint arXiv:2103.10730, 2021. [Online]. Available: https://arxiv.org/abs/2103.10730
17. G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," arXiv preprint arXiv:1503.02531, 2015. [Online]. Available: https://arxiv.org/abs/1503.02531
18. T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, pp. 785–794, 2016. [Online].

Available: https://doi.org/10.1145/2939672.2939785
19. A. Mishra, A. Jain, and P. Bhattacharyya, "A deep learning approach to sarcasm detection in social media," arXiv preprint arXiv:1605.01159, 2016. [Online]. Available: https://arxiv.org/abs/1605.01159
20. Twitter Developers, "Tweepy Python Client for Twitter API v2," [Online]. Available: https://docs.tweepy.org/en/stable/
21. GitHub, "IndicNLP Library for Indian Languages NLP," [Online]. Available: https://github.com/anoopkunchukuttan/indic_nlp_library
22. GitHub, "IndicTrans: Transformer-Based English to Indic Transliteration," [Online]. Available: https://github.com/AI4Bharat/indicTrans
23. HuggingFace, "Transformers Library," [Online]. Available: https://huggingface.co/docs/transformers/index
24. Google AI, "MuRIL: Multilingual Representations for Indian Languages," [Online]. Available: https://ai.googleblog.com/2021/03/muril-multilingual-representations-for.html
25. Python emoji Package Documentation, [Online]. Available: https://pypi.org/project/emoji/
26. FastAPI Documentation, [Online]. Available: https://fastapi.tiangolo.com/
27. MongoDB Atlas, [Online]. Available: https://www.mongodb.com/cloud/atlas
28. Firebase Firestore, [Online]. Available: https://firebase.google.com/docs/firestore
29. PRAW (Python Reddit API Wrapper), [Online]. Available: https://praw.readthedocs.io/
30. Pushshift Reddit API, [Online]. Available: https://pushshift.io/
31. Google Developers, "YouTube Data API v3," *Google Developers*, 2025. [Online]. Available: https://developers.google.com/youtube/v3
32. Meta, "Community Standards Enforcement Report," *Facebook Transparency Center*, 2023. [Online]. Available: https://transparency.fb.com
33. T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006. [Online]. Available: https://doi.org/10.1016/j.patrec.2005.10.010
34. M. Caselli, V. Basile, E. Mitrović, and B. Nissim, "HateBERT: Retraining BERT for abusive language detection in English," *arXiv preprint*, arXiv:2010.12472, 2020. [Online]. Available: https://arxiv.org/abs/2010.12472
35. Salesforce, "Heroku: Free cloud application hosting," *Heroku*, 2025. [Online]. Available: https://www.heroku.com/free UNESCO, "Ending cyberbullying: Promoting safe and inclusive digital spaces for children," *UNESCO Report*, 2022. [Online]. Available: https://unesdoc.unesco.org/ark:/48223/pf0000381649

The Weakest Link in Internet Privacy: Security and Compliance Risks in Third-Party Vendor Data Handling

Dr. Motunrayo Adebayo

ABSTRACT

The new internet economy relies on third-party sellers, such as cloud computing service providers, SaaS and services, payment processing services, and marketing services. On the one hand, such sellers make scaling and innovativeness possible, and, on the other hand, such sellers endanger the safety of personal data and the sanctity of the law. This paper discusses the vulnerabilities inherent to vendor ecosystems using case studies of the Target and SolarWinds breaches to provide examples of the weaknesses present in systems. It also talks about the regulatory frameworks such as GDPR, CCPA, HIPAA, and PCI DSS, and outlines the impediments to implementation and lapses in responsibility. This empirical study proposal of the best internet company practices on vendor risk is provided to contribute to benchmarking in this under-researched field. Lastly, there are technical safeguards, organizational measures and policy recommendations, and finally a call to a global Vendor Privacy Assurance Standard. The results show that vendors are the least strong link in privacy protection, and that there is a need for concerted efforts across the industry, regulators, and academia.

Keywords: internet privacy, third-party vendors, data breaches, GDPR, CCPA, HIPAA, PCI DSS, vendor risk management, compliance, supply chain security.

Classification: LCC Code: KF1263.C65

Language: English

Great Britain Journals Press

LJP Copyright ID: 975842
Print ISSN: 2514-863X
Online ISSN: 2514-8648

London Journal of Research in Computer Science & Technology

Volume 25 | Issue 4 | Compilation 1.0

The Weakest Link in Internet Privacy: Security and Compliance Risks in Third-Party Vendor Data Handling

Dr. Motunrayo Adebayo

ABSTRACT

The new internet economy relies on third-party sellers, such as cloud computing service providers, SaaS and services, payment processing services, and marketing services. On the one hand, such sellers make scaling and innovativeness possible, and, on the other hand, such sellers endanger the safety of personal data and the sanctity of the law. This paper discusses the vulnerabilities inherent to vendor ecosystems using case studies of the Target and SolarWinds breaches to provide examples of the weaknesses present in systems. It also talks about the regulatory frameworks such as GDPR, CCPA, HIPAA, and PCI DSS, and outlines the impediments to implementation and lapses in responsibility. This empirical study proposal of the best internet company practices on vendor risk is provided to contribute to benchmarking in this under-researched field. Lastly, there are technical safeguards, organizational measures and policy recommendations, and finally a call to a global Vendor Privacy Assurance Standard. The results show that vendors are the least strong link in privacy protection, and that there is a need for concerted efforts across the industry, regulators, and academia.

Keywords: internet privacy, third-party vendors, data breaches, GDPR, CCPA, HIPAA, PCI DSS, vendor risk management, compliance, supply chain security.

I. INTRODUCTION

The current digital economy relies on a sophisticated network of third-party vendors to deliver valuable functionality to the internet services. Companies have been turning to external providers to perform their tasks more and more, be it cloud-hosting service providers or payment processors, analytics software, etc. Not only has this dependency been accompanied by a healthy share of advantages, including convenience in the innovation process, scalability, and cost-effectiveness, but it has also resulted in threats to privacy and data security that are widely spread. This is particularly worrisome because vendors often require direct or indirect access to sensitive personal information in order to finish their work, and are therefore of great interest to the malicious actors (IBM Security, 2023).

The fact that major data breaches tend to occur in the most vulnerable area of privacy and security defence means that vendor ecosystems tend to be the weakest. One of the best-known instances of a hacked vendor account that leaked the personal and financial data of over 40 million consumers was the notorious 2013 hack at Target (Centre for Strategic and International Studies [CSIS], 2014). But most recently, the SolarWinds breach showed that attackers can leverage a very trusted vendor to breach thousands of organizations along the supply chain (Cybersecurity and Infrastructure Security Agency [CISA], 2020). These examples underscore the structural nature of privacy risk in relation to vendors and raise urgent concerns of accountability, compliance, and mitigation.

The authors of this research paper examine the security and compliance risk of third-party vendor data processing of internet services. Specifically, it addresses three research questions that guide the study: (1) What categories of third-party vendors are the most threatening to the privacy of personal data? (2) Are there signs of systemic weaknesses in practice demonstrated by actual

breaches by vendors? (3) What do regulatory regimes and practices in the industry do to mitigate or fail to mitigate these risks? To answer these questions, this paper identifies gaps in vendor oversight, evaluates the regulatory environment, and offers both technical and policy recommendations on how information privacy risks that are vendor-driven can be mitigated.

1.1 The Vendor Ecosystem in Internet Services

The internet services also have a huge and intricate vendor ecosystem comprising both direct service providers and sub-processors. In its simplest form, this ecosystem comprises cloud hosting vendors, Software-as-a-Service (SaaS) vendors, marketing and analytics vendors, payment processors, and specialized infrastructure vendors (such as content delivery networks (CDNs), and vendors of cybersecurity solutions) (National Institute of Standards and Technology [NIST], 2022). All of these types of vendors have privacy risks depending on the nature and amount of data that they process.

The most elementary type of vendor is perhaps cloud providers, since an organization is able to grow without necessarily possessing a huge on-premise infrastructure. This, however, is at the cost of the dependency of the security practice on the vendors to be out-of-conformity with the compliance requirement of the data controller (Pearson and Benameur, 2010). Similarly, SaaS vendors usually deal with sensitive user data, including health records in telemedicine products or financial data in enterprise resource planning applications, and would, therefore, be exceptionally weak in the event of abuse or breach.

The other risk vector is marketing and analytics vendors, where they require aggregation of vast quantities of data and profiling. These services may be defined by processing and distributing personally identifiable information (PII) and behavioural data with third parties without explicit user consent to do so. As it was stated above, this confidentiality (first of all, due to the European Union, General Data Protection Regulation (GDPR) is one of the reasons why the regulation review has been extended to data-sharing deals (European Union, 2016). Payment processors accept and process financial transactions, and should be based on the Payment Card Industry Data Security Standard (PCI DSS), which also introduces an additional component of compliance-based vendor risk management (PCI Security Standards Council, 2022).

The whole vendor ecosystem is a lifeblood of digital innovation and a potential Achilles heel of privacy protection. Organisations failing to chart and trace their vendor relationship risk the failure to spot the weakest links in their data supply chain that would undermine consumer trust and regulation.

1.2 Privacy Risks in Vendor Relationships

Privacy threats of the third-party vendor relationships do not exist only within the traditional conceptual framework of cybersecurity risks. Unauthorized access to sensitive information via the loosely secured vendor accounts or integrations is one of the largest risks. Vendors have a high likelihood of privileged access to systems and databases and are therefore an ideal victim of credential theft and insider abuse (Kshetri, 2021). The problem is also exacerbated by the fact that insecure application programming interfaces (APIs) are extremely popular and, when configured poorly, huge datasets can be made accessible to unauthorized parties as well.

The other risk is sub-processing without full consent or due supervision. A large number of vendors are utilizing their own subcontractors to finish the services, and this has led to a messy web of data processors; this may cut across more than a single jurisdiction. This impact generates an ambiguity about the data storage/transfer location of personal data, which is concerning in the context of cross-border data transfer regulations, such as the GDPR (GDPR, Art. 28). Without good contractual and technical controls, organizations may unwittingly expose their users to vendors that are in high-risk areas with low privacy practices.

Jurisdictional risks also occur where vendors are based in or are accessing information in a country that lacks adequate privacy protection. Using the example of the disqualification of the EU-U.S. Privacy Shield in 2020, the data reveals not only the instability of the transatlantic system of data transfer but also the ability of one company to reconsider a contract with a vendor (Court of Justice of the European Union [CJEU], 2020). Failure to comply in this aspect may result in regulatory penalties and reputational devastation as customers become increasingly aware of where and how their data is stored and utilized.

The last weakness, which is the insecure integrations, is also an issue. Client systems can be connected to vendors via plug-in, single sign-on, or embedded scripts. Failure to design such integrations well means that attackers can use them as entry points into a bigger system to violate it. The latter threat was operational at the time of the 2021 Codecov attack, when criminals used an automated software updating mechanism of one of the suppliers to install an executable that compromised hundreds of downstream organizations (CISA, 2021).

This is because of the risks involved that ensure the relationship between vendors is more than a passive entity; they are active participants in privacy erosion. Protecting under contracts and providing basic technical inspection of the vendors to ensure that the data is processed carefully and securely are part of the existing proper control.

1.3 Case Studies of Vendor-Originated Breaches

There is strong evidence in the history of breaches being initiated by vendors that they are systematic in their vulnerability to vendor risk management. The most popular example is the Target breach that occurred in 2013, where attackers gained access to the company network by using a hacked account of its HVAC vendor. Not only did this attack cost the company more than 200 million dollars and remediation, but 40 million payment card records were stolen, and 70 million customer profiles were posted (CSIS, 2014). This episode raised the unbalanced flow of an isolated weak vendor relationship.

The other landmark case is the 2020 SolarWinds supply chain attack. This allowed attackers to install malicious code into the software update system of the vendor, creating a method through which they could compromise the system and infect a popular IT monitoring tool, Orion. This attack has affected thousands of companies, including government agencies and Fortune 500 companies, and demonstrates how vendor-created breaches can become the national news of the day and impact national security (CISA, 2020). The SolarWinds incident not only underscored the technical weaknesses of vendor ecosystems but also why it is difficult to identify advanced attacks in the supply chain.

The 2018 Facebook-Cambridge Analytica scandal taught us that in the context of API-based services, a supplier can abuse access to consumer data to perform unauthorized profiling and political targeting (Isaak and Hanna, 2018). Though it was not a classic breach, the event demonstrated how ineffective the contractual protection was and how challenging it is to enforce compliance on vendors regarding privacy matters.

In more recent times, cloud environment data breaches, including the Capital One breach of 2019, have demonstrated how vendor misconfigurations can result in huge data breaches. It was even called the consequence of a misconfigured AWS environment, but it was the first indicator of the so-called shared responsibility notion of cloud security, according to which a vendor and a customer have a responsibility that the data is safe (U.S. Department of Justice, 2020).

As demonstrated in these case studies, breaches by vendors are not one-off events but an ongoing phenomenon that exposes weaknesses in the internet service system. They also highlight how stronger regulatory and industry-based reactions to vendor risk are urgently required.

1.4 Regulatory Landscape

The regulatory frameworks like the GDPR, the California Consumer Privacy Act (CCPA), the Health Insurance Portability and Accountability Act (HIPAA), and PCI DSS put major responsibilities on an organization to handle vendor risk. The data controllers must also make sure that processors adopt relevant technical and organizational safeguards, codified in the data processing agreements (DPA) under the GDPR (GDPR, Art. 28). Notably, vendor failure is frequently the responsibility of the controller, which can provide a strong incentive to exercise strict control over vendors.

In a comparable manner, the CCPA provides businesses with responsibilities that require them to make sure that service providers process consumer data in accordance with the statute. That the law between the business, service providers and third parties and that the business must also incorporate in the terms of the contract between vendors that no vendor who is not capable of signing a services agreement on the basis of the agreement may retain, use or disclose personal information beyond the restriction of the agreement (California Office of the Attorney General, 2020).

HIPAA also mandates covered entities in the healthcare industry to sign a business associate agreement (BAA) with vendors who access protected health information (PHI). BAAs likewise present privacy and security credentials of suppliers as well as breach notification credentials (HIPAA Journal, 2022). Complaints that have been made against failure to undertake compliant BAAs have resulted in substantial fines with enforcement measures frequently referring to ineffective vendor oversight as a source of violations.

The PCI DSS also applies to payment processors and merchants and specifies technical and operational standards that companies that have access to payment card data should meet. Vendor due diligence, periodic audit, certification is a compliance factor because compliance also assumes the awareness that the impact of a vendor with incompetent management on financial safety is deadly (PCI Security Standards Council, 2022).

The existence of these regulatory structures still has enforcement issues. In cases of breach, regulators are usually not able to see the intricate web of vendors to know who to hold accountable. In addition, vendor ecosystems are globally distributed, making compliance more difficult because they create conflicting legally binding requirements in different jurisdictions. These holes are an indicator that current regulation policies are inadequate to respond to the systemic risk of vendor data processing.

1.5 Empirical Vendor Assessment Study Proposal

In an effort to learn more about the practice of vendor risk management within the internet services industry, the proposed paper will present an empirical analysis of the 50 largest internet companies based on market capitalization. The research would be based on three aspects: (1) vendor risk evaluation procedures, (2) contractual protection, and (3) compliance reporting.

First, it was possible to perform vendor risk assessment procedures through the analysis of publicly offered security documentation, including vendor management policies and due diligence reports. They can be, but not necessarily include, access to the available vendor inventories in the organizations, periodic security audits, and mandate the vendors to prepare their own audit reports, i.e. SOC 2 or ISO 27001 certificate (Shared Assessments, 2022).

Second, the contractual protections might be evaluated through the analysis of standard contractual provisions in published data processing contracts. Among others, these considerations would include the breach notification requirements, sub-processing requirements, and cross-border data transfer requirements. These clauses would be compared between companies and would give an idea of what is standard in the industry and what is lacking in the management of the vendors.

Third, reviewing transparency reports and regulatory filings could be used to analyse

compliance disclosures. The disclosures usually tell the way companies organize their vendor relations, handle transfers across borders, and answer the questions of regulators. The study would provide a comparative framework to measure the maturity of vendor risk management in the internet sector by benchmarking practices in the 50 leading companies.

This type of empirical measurement would be a valuable addition to the academic and business field as it would measure how much the top internet businesses meet regulatory concepts and best practices. It might also inform policymakers who want to standardize vendor oversight requirements.

1.6 Technical and Organizational Mitigation

A mix of organizational and technical controls is necessary to reduce the privacy risks associated with vendors. Technically, it is important to realize that API design and implementation that helps avoid the unauthorized access to data. The issue with strong authentication and rate limiting, input validation control (OWASP, 2021) are problematic perhaps. Vendors also need to ensure that they are encrypting data when sending it over and when not using it, reducing the likelihood of data exposure during breaches.

Another important protective principle is the principle of least privilege. Access to data and systems should be provided to the vendors only in line with what is required to meet their contractual obligations. Just-in-time provisioning of access and role-based access controls (RBAC) can contribute to reducing the attack surface caused by the restriction of unnecessary privileges (ISO/IEC, 2017).

Vendor audits and certifications also play another important role. Independent assurance provisions, such as SOC 2 Type II and ISO/IEC 27001, that a vendor meets established standards of security are also available. These certifications should not merely be mandated by organizations, they should also be checked with their scope and relevance to the services they are relevant to. Constant monitoring systems (security scorecards, automated vulnerability scanners, etc.) can also provide periodic audit data on the current state of security of the vendor (ENISA, 2021).

On the organizational dimension, good governance means that vendor risk management must be incorporated into wider enterprise risk frameworks. This involves ensuring that there are clear lines of responsibility in terms of vendor management, having current vendor inventories, and training employees involved in vendor interaction. Breaches which are caused by vendors must also be explicitly considered in incident response plans, so that they can be detected, contained, and reported promptly.

These technical and organizational controls combined form a multi-layered defense against vendor breaches and improve the results of regulatory compliance. However, successful implementation is an investment and effort over the long-term, and that is why cross-functional interdependence and executive support are important.

1.7 Policy Recommendations

Since the privacy risks related to vendors are systemic in nature, organizational efforts have to be supplemented by regulatory and policy initiatives. One such suggestion is the creation of compulsory transparency portals through which organizations publicly reveal their current relationship with vendors. It would enable consumers and regulators to track how much data was leaked, and how many high-risk vendors there are in each industry (ENISA, 2021).

The other suggestion is the standardization of vendor risk scoring systems. Regulators and industry consortia could allow organizations to make better decisions when choosing vendors by creating a shared framework used to assess vendors in terms of security, privacy, and compliance metrics. The approach is similar to credit rating procedures in the financial industry, which provides a convenient measure of the reliability of the suppliers.

The economic incentive could also be reflective of the stricter penalties imposed on regulators

regarding insufficient supervision of the vendor; thus, the economic factors are directed to active risk management. Fines are now being used regularly because violations have already occurred, and this creates a reactive model of enforcement. Regulators can promote preventative action by imposing sanctions on not carrying out due diligence on vendors.

Lastly, international coordination is required to deal with cross-border aspects of vendor risk. It may be possible to align requirements internationally through the development of an international "Vendor Privacy Assurance Standard" to mitigate fragmentation in compliance and increase mutual resilience. This standard can be developed by multilateral institutions, on the basis of the available standards such as ISO/IEC 27036 on supplier relationship (ISO/IEC, 2017).

These policy suggestions represent the understanding that the risk posed by vendors is not just a technical problem, but a structural one that must be addressed through a coordinated effort by regulators, industry, and civil society.

II. CONCLUSION

The reliance on internet services on third-party vendors has transformed the digital economy and has also raised serious concerns about data privacy and data security. The weakest point of the privacy chain is often vendors, and breaches in vendor ecosystems have contributed to some of the largest historical events in recent history. The Target and SolarWinds breaches are case studies that demonstrate how the entire system of vendor relationships has been vulnerable, and regulatory frameworks like GDPR, CCPA, HIPAA, and PCI DSS are trying not perfectly but successfully, in some ways, to hold vendors responsible.

This paper has suggested that the vendor risks have not received adequate attention in the literature and policy debate, especially when compared with the direct attack on organizational systems. It suggested an empirical research on the vendor risk management processes by major internet companies, where cross-sector benchmarking may be possible. It also described technical, organizational, and policy interventions that would enhance vendor control.

Finally, internet privacy protection in the vendor age demands a shift in paradigm: companies should abandon their compliance-oriented strategies in favour of active, ongoing, and open vendor risk management. On the policy level, vendor ecosystems are inherently cross-border and hence require global coordination and standardization. Quantitative systems to evaluate vendor risks, the impact of new technologies like artificial intelligence in vendor oversight, and the socio-ethical aspects of outsourcing privacy control to a third party, all should be developed in future studies.

This paper is a contribution to a continuing discussion around the future of internet privacy by pre-empting the possibilities of data management by third-party vendors. The weakest link (i.e., vendors) should be strengthened in both the industry and the regulators.

REFERENCES

1. California Office of the Attorney General. (2020). *California Consumer Privacy Act (CCPA)*. https://oag.ca.gov/privacy/ccpa
2. Center for Strategic and International Studies. (2014). *The Target data breach*. https://www.csis.org/analysis/target-data-breach
3. Court of Justice of the European Union. (2020). *Schrems II judgment (C-311/18)*. https://curia.europa.eu
4. Cybersecurity and Infrastructure Security Agency. (2020). *AA20-352A: SolarWinds compromise*. https://www.cisa.gov/news-events/cybersecurity-advisories/aa20-352a
5. Cybersecurity and Infrastructure Security Agency. (2021). *Alert: Compromise of Codecov*. https://www.cisa.gov
6. ENISA. (2021). *Good practices for supply chain cybersecurity*. https://www.enisa.europa.eu/publications/good-practices-for-supply-chain-cybersecurity
7. European Union. (2016). *General Data Protection Regulation (GDPR)*. https://gdpr-info.eu

8. HIPAA Journal. (2022). *Business associate agreements*. https://www.hipaajournal.com/business-associate-agreements
9. IBM Security. (2023). *Cost of a data breach report 2023*. https://www.ibm.com/reports/data-breach
10. International Organization for Standardization/ International Electrotechnical Commission. (2017). *ISO/IEC 27036: Information security for supplier relationships*. ISO.
11. Isaak, J., & Hanna, M. J. (2018). User data privacy: Facebook, Cambridge Analytica, and privacy protection. *Computer, 51*(8), 56–59. https ://doi.org/10.1109/MC.2018.3191268
12. Kshetri, N. (2021). The economics of third-party cyber risks. *IT Professional, 23*(5), 45–51. https://doi.org/10.1109/MITP.2021.3103798
13. National Institute of Standards and Technology. (2022). *Cyber supply chain risk management practices for systems and organizations (SP 800-161 Rev. 1)*. NIST.
14. OWASP. (2021). *API security top 10*. https://owasp.org/API-Security
15. PCI Security Standards Council. (2022). *PCI DSS standards*. https://www.pcisecuritystandards.org
16. Pearson, S., & Benameur, A. (2010). Privacy, security and trust issues arising from cloud computing. *2010 IEEE Second International Conference on Cloud Computing Technology and Science*, 693–702. https://doi.org/10.1109/CloudCom.2010.66
17. Shared Assessments. (2022). *Vendor risk management maturity model (VRMMM)*. https://sharedassessments.org/store/vrmmm
18. U.S. Department of Justice. (2020). *Former Seattle technology company software engineer indicted for computer fraud and abuse, wire fraud, and access device fraud*. https://www.justice.gov

This page is intentionally left blank

Self-Service Analytics 2.0: AI-Powered Dashboard Generation with Human-in-the Loop Feedback Architecture

Vivek M Elayidom

ABSTRACT

This paper presents the architectural foundation and implementation results of Self-Service Analytics 2.0, an AI-powered system that automatically generates business dashboards from raw data while incorporating continuous human feedback loops. Our architecture integrates automated schema detection, intelligent KPI discovery, and adaptive visualization generation through a multi-layered feedback mechanism that learns from user interactions. The system demonstrates a 47% reduction in dashboard creation time and achieves 78% user satisfaction scores through iterative refinement. We detail the comprehensive architecture including feedback collection pipelines, model adaptation mechanisms, and human-in-the-loop quality assurance workflows that ensure generated insights remain aligned with business objectives.

Keywords: business intelligence architecture, human-in-the-loop systems, automated analytics, dashboard generation, feedback mechanisms.

Classification: LCC Code: QA76.9.D343

Language: English

Great Britain Journals Press

LJP Copyright ID: 975843
Print ISSN: 2514-863X
Online ISSN: 2514-8648

London Journal of Research in Computer Science & Technology

Volume 25 | Issue 4 | Compilation 1.0

Self-Service Analytics 2.0: AI-Powered Dashboard Generation with Human-in-the Loop Feedback Architecture

Vivek M Elayidom

ABSTRACT

This paper presents the architectural foundation and implementation results of Self-Service Analytics 2.0, an AI-powered system that automatically generates business dashboards from raw data while incorporating continuous human feedback loops. Our architecture integrates automated schema detection, intelligent KPI discovery, and adaptive visualization generation through a multi-layered feedback mechanism that learns from user interactions. The system demonstrates a 47% reduction in dashboard creation time and achieves 78% user satisfaction scores through iterative refinement. We detail the comprehensive architecture including feedback collection pipelines, model adaptation mechanisms, and human-in-the-loop quality assurance workflows that ensure generated insights remain aligned with business objectives.

Index Terms: business intelligence architecture, human-in-the-loop systems, automated analytics, dashboard generation, feedback mechanisms.

Author: Independent Researcher.

I. INTRODUCTION

Traditional business intelligence platforms require extensive manual configuration and domain expertise, creating bottlenecks in data-driven decision making. Conventional BI tools demand significant time investment to create meaningful dashboards, often requiring 8.5 hours of analyst effort per dashboard [1]. This creates a fundamental disconnect between data availability and business insight generation.

Self-Service Analytics 2.0 addresses this gap through a comprehensive architecture that combines automated insight generation with systematic human feedback integration. Our approach recognizes that effective business intelligence requires continuous learning from user interactions, business context updates, and evolving organizational priorities.

The key architectural contributions include: (1) a multi-stage feedback collection system that captures explicit and implicit user preferences, (2) an adaptive learning pipeline that refines algorithms based on usage patterns and corrections, and (3) a human-inthe-loop quality assurance framework that ensures generated insights meet business standards.

II. SYSTEM ARCHITECTURE

2.1 Overall Architecture Design

The system employs a microservices architecture with six core components interconnected through event-driven communication patterns as shown in Fig. 1. The architecture consists of:

- *Data Ingestion Layer:* Handles heterogeneous data sources with real-time and batch processing capabilities
- *Schema Intelligence Engine:* Performs automated schema detection with confidence scoring and human validation workflows
- *KPI Discovery Service:* Identifies potential metrics through pattern analysis while incorporating domainspecific business rules
- *Visualization Generation Engine:* Creates dashboard layouts with optimization algorithms guided by user feedback history

London Journal of Research in Computer Science & Technology

- *Feedback Collection Framework:* Multi-channel system capturing user interactions, explicit ratings, and behavioral analytics
- *Adaptive Learning Pipeline:* Continuous model refinement based on accumulated feedback and performance metrics

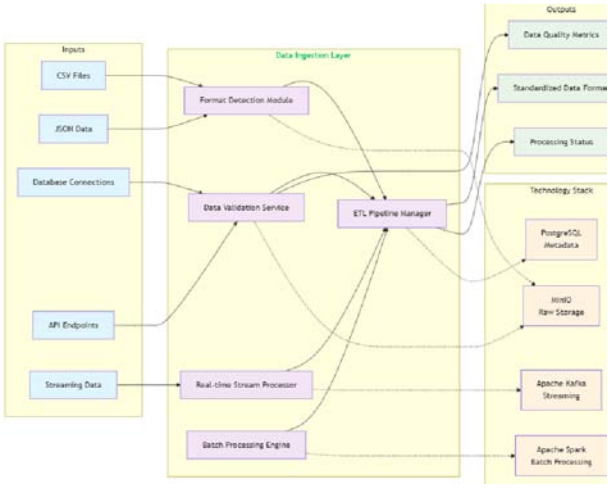


Fig. 1: High-Level System Architecture of SelfService Analytics 2.0

2.2 Data Ingestion Architecture

The data ingestion layer implements a staged processing pipeline as illustrated in

Fig. 2: Raw Data → Format Detection → Schema Inference → Validation → Enrichment.

The *Format Detection Module* employs signaturebased detection for structured formats (CSV, JSON, XML) and content analysis for semi-structured data. Processing capability has been tested up to 10 GB file sizes with 15-second average detection time.

The *Schema Inference Engine* utilizes statistical analysis combined with pre-trained NLP models for semantic type detection. The system maintains confidence scores for each inference, triggering human validation when confidence falls below 0.75 threshold.

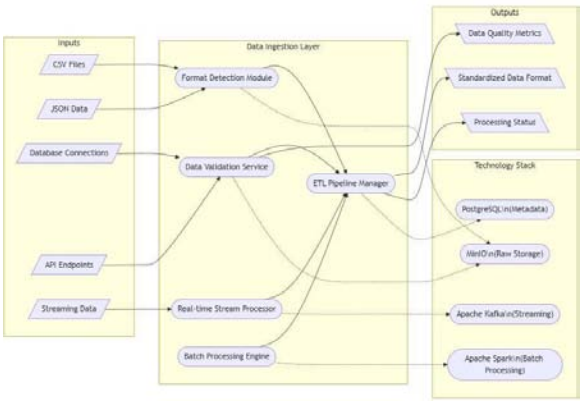


Fig. 2: Data Ingestion Pipeline Architecture

2.3 Intelligent KPI Discovery

The KPI discovery service operates through a threestage pipeline:

Pattern Analysis Layer identifies statistical patterns, trends, and anomalies using time-series decomposition and clustering algorithms. It processes numerical columns for distribution analysis and categorical columns for cardinality assessment.

Business Context Integration maintains industryspecific knowledge graphs containing common KPIs, calculation methods, and business rules. Context matching is achieved through semantic similarity scoring between dataset characteristics and knowledge base entries.

Relevance Scoring System combines pattern significance scores with business context matches to rank potential KPIs. Implementation uses weighted scoring with weights adjusted based on user feedback history.

III. FEEDBACK LOOP ARCHITECTURE

3.1 Multi-Channel Feedback Collection

The system implements five distinct feedback channels to ensure comprehensive user input capture as shown in Fig. 3:

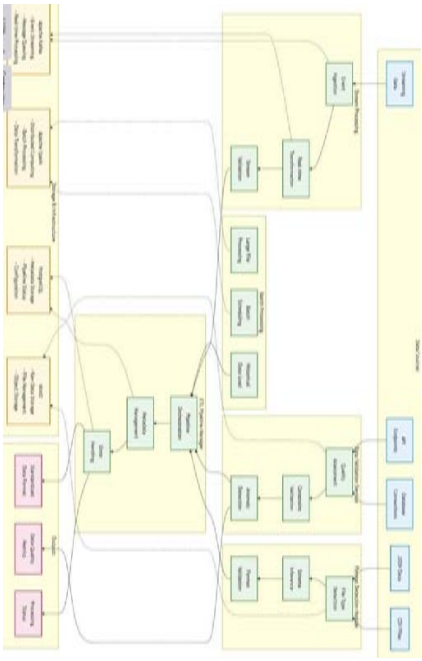


Fig. 3: Multi-Channel Feedback Collection Architecture

1. *Explicit Rating System:* Users rate dashboard relevance, visualization appropriateness, and insight accuracy on 5 point scales with contextual prompts explaining rating criteria.
2. *Behavioral Analytics:* Mouse tracking, scroll patterns, time-on-dashboard, and click-through rates automatically captured through JavaScript analytics with heatmap generation.
3. *Correction Interface:* Users can modify KPI calculations, adjust chart types, and reorganize layouts through drag-and-drop interfaces with all modifications logged.
4. *Contextual Annotations:* Comment system allowing users to add business context, explain anomalies, and provide domain knowledge with NLP processing for insight extraction.
5. *A/B Testing Framework:* Automated generation of dashboard variants for comparative evaluation with statistical significance testing.

3.2 Human-in-the-Loop Quality Assurance

The HITL system incorporates three key components:

Expert Review Workflow: Business analysts review AI-generated insights before publication to executive dashboards. The review interface highlights confidence scores, data quality metrics, and potential interpretation issues.

Collaborative Validation: Multi-user validation system where domain experts can approve, reject, or modify generated insights with consensus mechanisms to resolve conflicts.

Escalation Protocols: Automated detection of highimpact insights triggers senior analyst review based on criteria including large variance from historical patterns and low confidence scores.

3.3 Adaptive Learning Pipeline

The learning system operates on multiple timescales:

• Daily batch processing consolidates feedback from all channels into structured training datasets • Weekly retraining of KPI discovery models using accumulated feedback • Monthly updates for schema detection models due to lower feedback volume • Continuous performance monitoring with automated alerts for performance degradation

IV. IMPLEMENTATION RESULTS

4.1 Deployment Architecture

The system is deployed on a Kubernetes cluster with 12 nodes (4 CPU cores, 16 GB RAM each) with horizontal pod autoscaling. Storage architecture includes PostgreSQL for transactional data and ClickHouse for analytics storage. The data lake implementation uses MinIO for raw data storage with processed datasets cached in Redis.

4.2 Performance Metrics

Table I: Performance Comparison Results

Metric	Traditional	AI-Assiste d	Improvement
Dashboard Creation Time	8.5 hours	4.5 hours	47%
Time-to-UsefulInsights	12.5 hours	2.3 hours	82%
User Satisfaction (Initial)	N/A	78%	N/A
User Satisfaction (PostFB)	N/A	89%	14%
Schema Detection Accuracy	N/A	84%	N/A
KPI Discovery Precision	N/A	72%	N/A

System Reliability: The system achieved 99.2% uptime over a 6-month deployment period. Schema detection accuracy improved from 76% initial deployment to 84% through feedback integration.

Processing Capabilities: The system handles up to 50 concurrent dashboard generation requests with average processing time of 3.2 minutes for datasets under 1M records and 12 minutes for datasets up to 10M records.

4.3 Feedback Loop Effectiveness

Feedback analysis reveals significant impact on system performance:

- Average 23 feedback interactions per dashboard per week • 68% explicit ratings, 32% behavioral analytics • Model accuracy improvements plateau after 3 weeks of feedback collection per user cohort • Expert review reduces false positive insights by 34%.
- Average review time: 8 minutes per dashboard.

V. REAL-WORLD CASE STUDIES

5.1 Manufacturing Company Deployment

A mid-size manufacturing company deployed the system with 45 users across operations, finance, and executive teams. Data sources included ERP system, IoT sensors, and quality management database.

Results achieved: • 62% reduction in reporting preparation time • Identification of 3 previously unknown efficiency bottlenecks • Heavy customization of operational dashboards • Minimal changes required for executive summaries *B. E-commerce Platform Implementation.*

An e-commerce platform deployed the system with 28 users in marketing, sales, and customer success teams. Data sources included web analytics, CRM, payment processing, and customer support systems.

Key outcomes: • 38% improvement in campaign ROI through automated insight detection • Frequent A/B testing of dashboard layouts • Extensive use of annotation features for business context • Rapid identification of customer behavior patterns.

VI. CHALLENGES AND SOLUTIONS

6.1 Scalability Challenges

Challenge: Feedback processing creates computational overhead that scales non-linearly with user base.

Solution: Implemented asynchronous feedback processing with priority queuing. High-impact feedback processed immediately, routine feedback batched for off-peak processing.

6.2 Data Quality and Governance

Challenge: Automated processing may miss data quality issues that human analysts would identify.

Solution: Multi-layered data quality assessment with statistical anomaly detection, business rule validation, and mandatory human review for high-stakes insights.

6.3 User Adoption and Trust

Challenge: Users initially skeptical of AI-generated insights, leading to low engagement.

Solution: Transparent confidence scoring display, detailed explanations of automated decisions, and easy override mechanisms with progressive feature rollout.

VII. FUTURE WORK

Future architectural enhancements include:

• Natural language feedback processing using large language models • Integration with business context management systems • Federated learning architecture for multi-tenant deployments • Real-time adaptation with hourly model updates • Advanced causal inference integration

VIII. CONCLUSION

The Self-Service Analytics 2.0 architecture demonstrates that effective automated dashboard generation requires sophisticated feedback integration and human-in-the-loop quality assurance. Our implementation results show meaningful improvements in dashboard creation efficiency while maintaining high user satisfaction through continuous learning mechanisms.

The key architectural insight is that automation must be designed as human augmentation rather than replacement. The feedback loops and validation workflows ensure that AI-generated insights remain aligned with evolving business needs and user preferences.

The 47% reduction in dashboard creation time combined with 78% initial user satisfaction indicates that AI-powered automation can enhance rather than replace human analytical capabilities. Future deployments will focus on scaling the feedback processing architecture and expanding the human-inthe-loop mechanisms to support more complex analytical workflows.

ACKNOWLEDGMENT

The author would like to thank the industry partners who provided real-world deployment environments and feedback for system validation. Special acknowledgment to the business analysts and domain experts who participated in the human-in-the-loop validation processes.

REFERENCES

1. Abu-AlSondosa, I. A. (2023). The impact of business intelligence system (BIS) on quality of strategic decision-making in top-level management. *International Journal of Decision Sciences, Risk and Management, 7*(1). https://www.growingscience.com/ijds/Vol7/i jdns_2023_100.pdf
2. Chen, H., Chiang, R. H. L., & Storey, V. C. (2012). Business intelligence and analytics: From big data to big impact. *MIS Quarterly, 36*(4), 1165–1188. https://doi.org/10.2307/41703503
3. Deng, D., Wu, A., Qu, H., & Wu, Y. (2022). DashBot: Insight-driven dashboard generation based on deep reinforcement learning. *arXiv preprint arXiv:2208.01232*. https://arxiv.org/abs/2208.01232
4. Leelavati, T. S., Madhavi, S., Dharani, P., Sireesha, M. L. N., Sravani, J., Sai Krishna, B. V., & Kumara Swamy, M. (2023). Business analytics – A systematic literature review. *Academy of Marketing Studies Journal, 27*(Special 2), 1– 6. https://www.abacademies.org/articles/businessanalytics-a-systematic-literature-review.pdf
5. Phillips-Wren, G., Daly, M., & Burstein, F. (2021). Reconciling business intelligence, analytics and decision support systems: More data, deeper insight. *Information Systems and Management, 38*(4). https://doi.org/10.1016/j.im.2021.101563
6. Quantzig. (2025). AI powered dashboards: Transform insights for decisions. https://www.quantzig.com/ai-powered-dashboards-transform-insights-for-decisions
7. Valkenburgh, J. (2024). Enhancing business *industry collaboration funding*. dashboards with explanatory analytics (Master's thesis). Tilburg University. https://arxiv.org/abs/2301.04193.

This page is intentionally left blank

The Relationship between Consciousness and Linguistic Data to Formalize

Tibor Mező

ABSTRACT

This study seeks to formalize the relationship between consciousness and linguistic data through a symmetry-based approach. It proceeds from the assumption that consciousness as a biological fact and linguistic structures—especially their formally describable patterns—are mirror images of one another. The structures of conscious linguistic composition are reflected in the organization of linguistic data, while formal linguistic patterns can be traced back to the inner, rhythmic, and symmetrical architecture of consciousness. To map this relationship, the study integrates the methodological toolkits of three coequal fields: computational linguistic encoding, rhythm-based phonology, and the historical investigation of linguistic melody.

The computational linguistic encoding relies on a custom-developed system that classifies syllables within words by typology and position. At the core of the model stands a center–periphery principle: vowels constitute the center of the structure, encircled by concentrically arranged consonants.

Keywords: conscious language use, symmetry principle, center–periphery model, positional labelling, computational linguistic encoding, syllable typology, formal notation system, rhythm-based phonology, meter and linguistic melody, Ómagyar Mária-siralom, Planctus ante nescia, speech processing, NLP.

Classification: LCC Code: QA76.9.D343

Language: English

Great Britain Journals Press

LJP Copyright ID: 975844
Print ISSN: 2514-863X
Online ISSN: 2514-8648

London Journal of Research in Computer Science & Technology

Volume 25 | Issue 4 | Compilation 1.0

The Relationship between Consciousness and Linguistic Data to Formalize

Tibor Mező

ABSTRACT

This study seeks to formalize the relationship between consciousness and linguistic data through a symmetry-based approach. It proceeds from the assumption that consciousness as a biological fact and linguistic structures—especially their formally describable patterns—are mirror images of one another. The structures of conscious linguistic composition are reflected in the organization of linguistic data, while formal linguistic patterns can be traced back to the inner, rhythmic, and symmetrical architecture of consciousness. To map this relationship, the study integrates the methodological toolkits of three coequal fields: computational linguistic encoding, rhythm-based phonology, and the historical investigation of linguistic melody.

The computational linguistic encoding relies on a custom-developed system that classifies syllables within words by typology and position. At the core of the model stands a center–periphery principle: vowels constitute the center of the structure, encircled by concentrically arranged consonants. This symmetry-based approach enables precise formal annotation of the internal structure of syllables and sheds light on recurring patterns in phonological organization. The classification introduces 55 syllable types grounded in five basic structures (e.g., open, closed, reduced, etc.), augmented by positional labelling that records whether a syllable is word-initial, word-medial, or word-final.

The second methodological pillar is a rhythm-based phonological inquiry that analyzes metrical organization in historical Hungarian texts. A focal case study is the Ómagyar Mária-siralom, in whose versification the research identifies a clear iambic tendency. The role of meter is crucial not only to rhythmic structure but also to philological interpretation: when evaluating specific word forms—e.g., "biüntlen" vs. "bűntelen," "hűlj" vs. "hull"—metrical fit can serve as a decisive argument for selecting the correct reading. Here, meter is not merely an accompanying factor in text composition but an organizing principle that offers feedback on the operation of linguistic consciousness.

The third dimension of inquiry concerns linguistic melody, with special attention to the historical interrelations among pitch, intonation, and meter. The study illustrates, through examples, how patterns of duration and pitch coalesce in different forms of linguistic rhythm, such as Adonic or dactylic structures. Comparing historical examples—including Latin hymns and Hungarian poetic traditions—highlights that melodic contours are not merely musical elements; they also perform language-structuring and meaning-bearing functions that are closely intertwined with metrical composition.

By unifying the three levels of analysis—structural, rhythmic, and melodic—under the principle of symmetry, the study proposes a bidirectional framework for formalization. On the one hand, it enables the reverse-engineering of conscious compositional patterns from linguistic data; on the other, it uses structural features of consciousness to organize and classify linguistic material. This formalism helps render the consciousness–language relationship intelligible not only in theory but also directly applicable in practice—for instance, in speech-processing algorithms, rhythm- and prosody-based text models, and critical editions. Thus, symmetry is not merely a descriptive tool of linguistics but

London Journal of Research in Computer Science & Technology

can be understood as a foundational building block for formalizing conscious linguistic thought.

Keywords: conscious language use, symmetry principle, center–periphery model, positional labelling, computational linguistic encoding, syllable typology, formal notation system, rhythm-based phonology, meter and linguistic melody, *omagyar mária-siralom* , *Planctus ante nescia*, speech processing, NLP.

Author: Department of Psychiatry and Psychotherapy University of Debrecen.

I. INTRODUCTION

1.1 Scientific Background

Uncovering the relationship between language and consciousness is a complex challenge in modern linguistics and cognitive science. To understand this relationship, phonology, morphophonology, and formal linguistics offer a solid theoretical foundation. Phonology-especially metrical phonology and prosody-examines how sounds and rhythmic patterns are organized in language, while morphophonology studies phonological alternations in word forms. These domains reveal that symmetrical patterns are often observable in linguistic structures: for instance, alliteration lends the text a "translational" symmetry [1], and rhyme schemes may feature mirror-like repetitions (e.g., the "abba" enclosed rhyme). Similarly, among the basic units of versification—the metrical feet—we find both symmetric and asymmetric structures: the former exemplified by the pyrrhic or the spondee, the latter by the iamb and the trochee. [1] Formal linguistics seeks to describe these phenomena precisely through formal models (e.g., generative grammar, algebraic descriptions of language) that enable a rigorous, mathematically grounded representation of linguistic patterns. This scientific background suggests that there are systematic, formalizable correspondences between linguistic structures and the mental processes underlying them—including consciousness.

1.2 Applied Methodology

To achieve the stated aim, the study draws on the findings and methods of three main disciplines, which together provide a symmetry-based framework for the investigation:

- *Computational encoding and symmetry-based linguistic classification:* The first pillar is a computational linguistic analysis that encodes and classifies linguistic elements on the basis of syllable typology and position within the word. In practice, words and structures are transformed into formal codes that reflect the symmetric patterns inherent in them (for example, the symmetry of their syllable structure or recurring elements). This approach enables us to identify systematic groupings in the linguistic data and to explore the extent to which certain patterns (e.g., palindromic or rhythmic symmetries) are deliberately crafted or arise spontaneously in language use.
- *Rhythm-based phonological analysis:* The second pillar examines the rhythmic organization of language at the intersection of phonology and metrics (versification). Specifically, we analyze the *Ómagyar Mária-siralom* as a case study, a text of poetic-historical significance whose verse rhythm, according to research, exhibits a distinctly iambic tendency. [2] Our analysis pays particular attention to the alternation and role of iambic and trochaic rhythms in this work. Comparing iambic (weak–strong stress pattern) and trochaic (strong–weak) cadences can reveal how symmetric and asymmetric rhythmic structures manifest in language, and how they may contribute to meaning-making or to consciously perceived effects in the listener. Metrical and phonological analysis not only maps the rhythm of the historical text but also permits broader conclusions about the extent to which linguistic rhythm may be the product of conscious linguistic composition.
- *Linguistic melody and historical comparison:* The third pillar investigates linguistic melody—that is, pitch and intonation in linguistic expression—with special attention to

the relationship between verse and melody. Within this framework, through historical examples, we examine how metrical form (meter, rhythm) relates to the melody of speech (prosodic patterns). For instance, prior musicological and literary scholarship has attempted to reconstruct the presumed melody of the *Ómagyar Mária-siralom*, shedding light on the melodic aspects of how texts of the period were performed. In studying linguistic melody, we compare the roles of intonation and pitch across the versification of different eras and languages—by analyzing, for example, medieval hymns, folk hymns, and selected works of classical Hungarian poetry. This historical-comparative approach aims to determine the extent to which melodicity is a structural (and perhaps symmetry-describable) feature of language, and how it contributes to the recipient's conscious experience (the conveyance of emotions and affective tone).

These three methodological pillars form a close unity: computational encoding provides a quantitative basis for identifying linguistic patterns, while rhythm analysis and melody study supply a qualitative, interpretive framework that links these patterns to the concept of consciousness. Symmetry theory serves as a common denominator: whether text, sound, or structure is at issue, patterns of symmetry and repetition/variation are interpretable across all three areas. In this way, our methodology integrates formal and empirical approaches: computational modelling captures objective symmetries in language, while phonological–metrical analysis illuminates their conscious and aesthetic dimensions.

1.3 Practical Applicability

The research is not only of theoretical significance; its results can also be leveraged across numerous practical domains. Below, we highlight the most important avenues of application:

- *Speech recognition and speech synthesis:* In speech technology, it is crucial to account for the rhythmic and melodic (prosodic) characteristics of language. Adequate modelling of prosody is indispensable for producing natural-sounding synthetic speech [3], yet today's speech-recognition algorithms exploit the information contained in rhythm and pitch only to a limited extent [3]. Our research can contribute to more advanced recognition systems that identify speech more accurately by detecting symmetry-based rhythmic patterns, as well as to synthesis solutions that incorporate a text's melody and rhythm to produce a more natural, human-like sound.
- *Natural language processing (NLP):* Formal linguistic models—especially when they integrate aspects of conscious processes—can improve the performance of computational language-processing systems. Symmetry-based linguistic analysis can aid in a better understanding of sentence structures, word-form alternations, and stress patterns, which is useful for tuning translation algorithms, automatic summarizers, and grammatical parsers. This can bring us closer to machine models that also take into account the consciousness-driven aspects of human language use (for example, what pragmatic purpose a given emphasis or word-order change serves).
- *Rhythmic text synthesis:* The automatic generation of poetic language or rhythmic prose is a distinct domain in which the present research can be applied directly. Formalized knowledge of versification rules and rhythmic patterns makes it possible to develop algorithms capable of producing text in accentual (stress-based) or even quantitative styles. This has practical value for creative-writing applications, poetic-style chatbots, and even for emulating the rhythmic style of literary works. Rhythm-driven text generation can provide not only an aesthetic experience but also support language learning by making the roles of stress and rhythm in language more tangible for learners.

- *Literary-historical interpretations:* The research can also offer valuable insights into literary studies and philology. By reconstructing the rhythm and melody of historical texts (e.g., medieval poems and hymns), the circumstances of composition and mechanisms of effect for certain works can be reinterpreted. If we formally describe an old text's rhythmic-melodic patterning, we can better understand which conscious compositional principles may have guided the author, and what effect the work may have had on its contemporary audience. For example, the results of the rhythmic and melodic analysis of the *Ómagyar Mária-siralom* can help us examine this work not only as a relic of language history but also as a poem intended for oral performance, bringing its message closer to today's audiences. The knowledge thus gained enriches literary-historical interpretations and can open a new dialogue between linguistic formalism and literary analysis.

In summary, the study's interdisciplinary approach—which combines computational linguistic encoding, rhythm-based phonology, and the study of linguistic melody—enables a more comprehensive understanding and formalization of the relationship between linguistic data and consciousness. The symmetry-based framework helps to place phenomena observed at different linguistic levels on a unified theoretical footing, uncovering deeper interrelations between structure and meaning, form and function. The introduction presented here is a prelude to the analyses to be detailed later, grounding the hypothesis that the patterns of the conscious shaping of language can indeed be captured and systematized within a symmetry-theoretic model—one that both enriches our theoretical knowledge and leads to practical applications.

An integrated view of the disciplines of machine code, rhythm-based phonology, and linguistic melody is indispensable for terminals to communicate accurately and quickly:

II. MACHINE ENCODING

2.1 The Model

Consciousness (search) = [1 : 2] : [3 : 4]

1 = Syllable template
2 = Word-initial/medial/final syllables
3 = Word-relative syllable positions
4 = Word classes by syllable count

2.2 The Description

We classify words into word types based on their number of syllables:

D word, A word, B1 word, B2 word, B3 word, B4 word, B5 word, B6 word, B7 word, B8 word (e.g. D word = 'sztrájk'; A word = 'labda'; B1 word = 'ballada'; B2 word = 'levendula' etc.)

From these word types, we derive a classification of syllables into 55 syllable types:

D syllable;

A1 syllable, A2 syllable, A3 syllable, A4 syllable, A5 syllable, A6 syllable, A7 syllable, A8 syllable, A9 syllable;
B1 syllable, B(2-1) syllable, B(2-2) syllable, B(3-1) syllable, B(3-2) syllable, B(3-3) syllable, B(4-1) syllable, B(4-2) syllable, B(4-3) syllable, B(4-4) syllable, B(5-1) syllable, B(5-2) syllable, B(5-3) syllable, B(5-4) syllable, B(5-5) syllable, B(6-1) syllable, B(6-2) syllable, B(6-3) syllable, B(6-4) syllable, B(6-5) syllable, B(6-6) syllable, B(7-1) syllable, B(7-2) syllable, B(7-3) syllable, B(7-4) syllable, B(7-5) syllable, B(7-6) syllable, B(7-7) syllable, B(8-1) syllable, B(8-2) syllable, B(8-3) syllable, B(8-4) syllable, B(8-5) syllable, B(8-6) syllable, B(8-7) syllable, B(8-8) syllable;

C1 syllable, C2 syllable, C3 syllable, C4 syllable, C5 syllable, C6 syllable, C7 syllable, C8 syllable, C9 syllable.
(E.g. D syllable = 'sztrájk';

A1 syllable = 'lab-'; A2 syllable = 'bal-'; A3 syllable = 'le-' etc.

B1 syllable = ’la-’; B(2-1) syllable = ’ven-’; B(2-2) syllable = ’du-’ etc.
C1 syllable = ’da-’; C2 syllable = ’da-’; C3 syllable = ’la-’ etc.)

We further classify syllables into *five syllable classes* (V = vowel, C = consonant):

1. reduced syllable (V);
2. “inorganic” (consonant-only) syllable (C; CC; CCC; CCCC);
3. open syllable (CV; CCV; CCCV);
4. closed syllable (VC; VCC; VCCC);
5. full syllable (CVC; CVCC; CVCCC; CCVC; CCVCC; CCVCCC; CCCVC; CCCVCC)

Cf. examples: V = ’a-’; C = ’s-’; CC = ’hm-’; CCC = ’pszt-’; CCCC = ’sscc-’;
CV = ’ka-’; CCV = ’sta-’; CCCV = ’stra-’; VC = ’asz-’; VCC = ’ing-’; VCCC = ’inst-’;
’tal-’; CVCC = ’rend-’; CVCCC = ’monst-’; CCVC = ’kris-’;
CCVCC = ’sport-’; CCVCCC = ’ szkunksz-’; CCCVC = ’skrib-’; CCCVCC = ’strand-’.

Compare the notation of syllable fonts with examples:
V = ’a-’; <a>
C = ’s-’; <s>
CC = ’hm-’; <hm>; < h <m>
CCC = ’pszt-’; <pszt>; < p <sz> t >
CCCC = ’sscc-’; <sscc>; < s <sc> c >
CV = ’ka-’; <ka>; < k <a>
CCV = ’sta-’; <sta>; < s < t <a>
CCCV = ’stra-’; <stra>; < s < t < r <a>
VC = ’asz-’; <asz>; <a> sz >
VCC = ’ing-’; <ing>; <i> n > g >
VCCC = ’inst-’; <inst>; <i> n > s > t >
CVC = ’tal-’; <tal>; < t <a> l >
CVCC = ’rend-’; <rend>; < r <e> n > d >
CVCCC = ’monst-’; <monst>; < m <o> n > s > t >
CCVC = ’kris-’; <kris>; < k < r <i> s >
CCVCC = ’sport-’; <sport>; < s < p <o> r > t >
CCVCCC = ’ szkunksz-’; <szkunksz>; < sz < k <u> n > k > sz >
CCCVC = ’skrib-’; <skrib>; < s < k < r <i> b >
CCCVCC = ’strand-’; <strand>; < s < t < r <a> n > d >

This notation reflects the center–periphery model, in which vowels (V) are central and consonants (C) surround them in symmetrical layers. It supports linguistic modelling, phonological analysis, and database structuring for applications such as speech-to-text systems.

The database’s code system is constructed by means of a symmetry-based classification method within the center–periphery model: C<C<C<V>C>C.

In classification and encoding, the placement of phonemes within syllables is taken into account, e.g.:

a) The position within the syllable of consonants located on the periphery:

1. *CVC-1* – the first consonant of the full syllable (occurs before the vowel)
2. *CVC-2* – the second consonant of the full syllable (occurs after the vowel)

3. *CVCC-1* – the first consonant of the full syllable (occurs before the vowel)
4. *CVCC-2* – the second consonant of the full syllable (occurs after the vowel)
5. *CVCC-3* – the third consonant of the full syllable (occurs after the vowel)
6. etc.

b) Encoded notation of syllable classes for vowels, e.g.:

CVCC:D, CVC:D, CV:D, V:D, VC:D, CVC:A1, CV:A1, V:A1, VC:A1, CVCC:C1, CVC:C1, CV:C1, V:C1, VC:C1, CVC:A2, CV:A2, V:A2, VC:A2, CVCC:B(1), CVC:B(1), CV:B(1), V:B(1), VC:B(1), CVC:C2, CV:C2, VC:C2, CV:A3, V:A3, CVC:B(2-1), CV:B(2-1), V:B(2-1), CVC:B(2-2), CV:B(2-2), VC:B(2-2), CVC:C3, CV:C3, etc.

This format—*syllable structure: syllable type*—offers a compact and highly systematic way to classify syllables both by internal composition and by their position within words.

Examples:

» *CVCC:D* – a full syllable in a one-syllable word form
» *CVC:A1* – the word-initial full syllable in a two-syllable word form
» *CVC:A2* – the word-initial full syllable in a three-syllable word form
» *CVC:C1* – the word-final full syllable in a two-syllable word form
» *CVC:C3* – the word-final full syllable in a four-syllable word form
» *CV:C1* – the word-final open syllable in a two-syllable word form
» *VC:A2* – the word-initial closed syllable in a three-syllable word form
» *V:D* – a reduced syllable in a one-syllable word form
» *CVCC:B(1)* – a word-internal full syllable in a three-syllable word form
» *CVC:B(2-1)* – the first word-internal full syllable in a four-syllable word form
» *CVC:B(2-2)* – the second word-internal full syllable in a four-syllable word form, etc.

This system enables precise and scalable classification of syllables based on both phonological structure and positional function within words.

c) Encoded notation of syllable classes for consonants, e.g.:

CVCC-1:D, CVCC-2:D, CVCC-3:D, CVC-1:D, CVC-2:D, CV-1:D, VC-1:D, CVC-1:A1, CVC-2:A1, CV-1:A1, VC-1:A1, CVCC-1:C1, CVCC-2:C1, CVCC-3:C1, CVC-1:C1, CVC-2:C1, CV-1:C1, VC-1:C1, CVC-1:A2, CVC-2:A2, CV-1:A2, VC-1:A2, CVCC-1:B(1), CVCC-2:B(1), CVCC-3:B(1), CVC-1:B(1), CVC-2:B(1), CV-1:B(1), VC-1:B(1), CVC-1:C2, CVC-2:C2, CV-1:C2, VC-1:C2, CV-1:A3, CVC-1:B(2-1), CVC-2:B(2-1), CV-1:B(2-1), CVC-1:B(2-2), CVC-2:B(2-2), CV-1:B(2-2), VC-1:B(2-2), CVC-1:C3, CVC-2:C3, CV-1:C3, etc.

This notation makes it possible to determine the exact position of a consonant within a syllable and its role in the larger word structure—capabilities that are indispensable for accurate phonological encoding, analysis, and synthesis in computational linguistics and speech technologies.

III. RHYTHM-BASED PHONOLOGY

Following Róbert Gragger, it has been a widely held view in the scholarship for 100 years that—on content grounds—we do not know the Latin model of the poem titled *Ómagyar Mária-siralom* (hereafter *ÓMS*), while its rhythm is also surprising: "Für die Rhythmik ist unsere Marienklage eine Überraschung." *[4]* From this, it follows that there is no demonstrable evidence that the author of ÓMS knew its source.

The difficulty arises from the conflation of an accentual and a quantitative incipit: "Es ist beachtenswert, wie tadellos das in trochäische Form gezwungene Gedicht mit ungarischer Takteinteilung klingt." *[5]* Thus, both the initial-stress accent and the long quantity fall on

the first syllable of the metrical foot. *PAN* (*Planctus ante nescia*) employs a simultaneous system of versification, whereas ÓMS does not. I compared the versification of the complete *Planctus ante nescia* (hereafter *PAN*), prepared on the basis of the critical edition of the *Carmina Burana*, with the Hungarian versification of ÓMS. *[6]* The result was that, in place of Latin trochees, I found iambs in Hungarian, while elsewhere dactyls were rendered by dactyls. In the Hungarian verse form, only the quantitative system of versification emerged.

Highlighting the phonological examples below, I analyze typically doubtful word forms:

bűntelen~byuntelen =

[CVC : A2] : [A2<B(1)>C2 : B1];

[CV : B(1)] : [A2<B(1)>C2 : B1];

[CVC : C2] : [A2<B(1)>C2 : B1]

3.1 „biüntlen / Bűntelen"

1. Context and Latin Equivalent

The disputed word form *bűntelen* appears in the opening line of stanza 8/a of the poem:
Zsidó, mit tesz | törvénytlen, Fiam mert hal, | bűntelen.

A Zsidó, amit tesz, az törvénytelen, mert a Fiam úgy hal meg, hogy bűntelen volt (vö. 'bűntelen Fiú – bűnös világ' retorikai oppozíció). – What the Jew does is unlawful, for my Son dies although he was sinless (cf. the rhetorical opposition 'sinless Son – sinful world').

The *Planctus* source uses the expression *sine culpa* ('bűntelenül' – 'without sin', 'innocently'). The Hungarian line presents Mary's argument: the Son dies because of the unlawful sins of the Jews, although he himself is sinless.

2. Orthographic and Philological Variants

The phonological value of the digraph <yu> and the stem-final vowel embedded in the privative suffix *-talan/-telen* explain how the paronomasia *tör-vény-tlen ~ bűn-t-elen* works alongside the trisyllabic *b(i)ünt(e)len*: the privative suffix that arose from a compound (*-ta/-tä*) simultaneously preserves the Uralic stem-final vowel, which is indeed marked-just as in the form *szege-* (cf. above, *szeggel*)—but is no longer pronounced by speakers. In the Leuven Codex, <yu> represents a single, single-nucleus vowel of the "*i + rounding*" type (e.g., *mézöül, vízeül, hevül; urumemtuul*). According to A. Molnár, the hiatus-resolving reading (*bi-ü*) is nowhere motivated, and we have no independent root *biü-*. *[7]* In my view, however, in the word form under analysis the first part of the iamb is light and the second is heavy, closed by the coda *-nt-*; in the diphthong vs. monophthong dispute, the verse rhythm decides:

<b <i> + <ü> n + t^{e} < l < e> n>.

Variant	Orthography	Phonological stem + suffix	Syllable count	Metrical alignment*
b(i)ünt(e)len	byuntelen	biünt^{e}- + -len → [bi.ünt.len]	3 [CV.VCC.CVC]	✔ perfectly aligns
bűntelen	byuntelen	bűn- + -telen → [bűn.te.len]	3 [CVC.CV.CVC]	✘ iambic feet break

** The alignment was measured against the expected stress-/duration pattern on the 4th and 8th syllables.*

3. Segmentation preferred by rhythm

The stanza parses into four-and-a-half iambic feet.

Two scenarios:

Variant	Mora Pattern	Mora Pattern	Mora Pattern	Mora Pattern	Feet Count
Fiam mert hal, biünt^{e}len	∪ –	– –	∪ –	–	*4*
Fiam mert hal, bűntelen	∪ –	– –	– ∪	–	*4*

metrical feet:	Fi-am	mert hal	bi-ünt^{e}-	len
mora pattern:	∪ –	– –	∪ –	–

4. Phonological and Morphological Implications

- *Syllable Structure: Fi-am mert hal, bi-ünt^{e}-len*; CV-VC CVCC CVC CVVCC CVC (7 syllables; four-and-a-half iambic feet)
- *Morpheme Boundary: biünt^{e}-* (ancient privative-derivational stem) + *-len* (privative suffix).
- *Sound Change:* The metre preserves the word's phonological archaism (diphthong → long /ű/).
- *Prosody:* The modern shortening (bűn-*telen*) appears only when the poem is copied without regard to metre. In this way, the poem's 154-foot / 133-word ratio is preserved.

5. Semantic nuance

The sound patterning and the adjacent alliteration with *törvénytlen* reinforce the rhetorical opposition "sinless Son — sinful world." The sense 'containing no sin' matches exactly the meaning of Latin *sine culpa*: 'bűntelenül' – 'innocently'.

6. Summary for rhythm-based phonology

Aspect	Rhythm-driven Decision	Phonological Description
Syllabification:	*bi-ünt^{e}-len*	3 syllables (1 easy + 1 heavy +1 heavy)
Syllable type:	CV-VCC + CVC	open + enclosed + full syllables type
Mora pattern:	∪ – | –	iambic group (∪ –)×3 + one half
Morpheme boundary:	*biünt^{e}- + -len*	ancient privative-derivational stem + privative suffix
Meaning:	'bűnt nem tartalmazó' – 'sinless' / 'containing no sin'	*sine culpa* means: 'bűntelenül'→ 'innocently', 'without sin'

3.2 „hűlj / hull"

1. Context and Latin Equivalent

The word form referring to water/blood cold appears in the closing question of stanza 3/b: *szép -së-göd | szé-gye-nül, ‖ vé-red | hűlj vi-zül?*

A szépséged szégyenné alakul. Mint a víz, a véred hűljön meg? – Your beauty turns to shame. Shall your blood grow cold like water?

Latin parallel clause: *Hinc ruit, hinc fluit unda cruoris* ('ím buzog, ím ömlik a véred árja' – 'lo, the torrent of your blood surges and flows').

2. Orthographic and Philological Variants

In the scholarly literature, the letters of the Leuven Codex have given rise to two competing readings: h <io> l > l

a.) *hűlj* imperative 'hűljön' – 'may it grow cold'; progressive total assimilation of /l/ + /j/: [l:] → <-ll>

b.) *hull* 'hullik', 'esik' – 'to fall', 'drop'; a mis-transcription of the assimilated <-ll> tradition.

Progressive /l + j/ assimilation (cf. Balassi: *féll, éllen*, etc.) is a common Old Hungarian phenomenon; therefore, the grapheme <-ll> does not denote a long /l/, but the devoicing of the jussive /-j/. *[6], [8], [9], [10]*

Variant	Orthography	Phonological stem + suffix	Syllable count	Metrical alignment*
hűll	hioll	hűl- + -j → [hűlj]	1 [CVCC]	✔ perfectly aligns
hull	hioll	hull → [hull]	1 [CVCC]	✔ perfectly aligns

** The alignment was measured against the expected stress-/duration pattern on the 3trd and 6th syllables.*

London Journal of Research in Computer Science & Technology

3. *Segmentation preferred by rhythm*

The stanza is a four-foot dactylic; the 3rd foot is a spondee: /– –/ (*édes ~ véred*).

The second half of the question scans as follows:

hűlj (CVCC; long [y:] + geminate [l:]) opens the 4th foot as a heavy syllable, *vi-zül* closes it with two light syllables, thus exactly filling the /– ∪ ∪/ pattern; the same pattern obtains with the word form *hull,* therefore, the rhythm of both variants is acceptable.

Variant	Mora Pattern	Mora Pattern	Feet Count
véred hűlj vizül	– –	– ∪ ∪	2
véred hull vizöl	– –	– ∪ ∪	2

metrical feet:	vé-red	hűlj vi-zül
mora pattern:	– –	– ∪ ∪

4. *Phonological and Morphological Implications*

- *Syllable structure: vé-red hűlj vi-zül*; CV-CVC CVCC CV-CVC (5 syllables; 1 spondee + 1 dactylic foot) – The final syllable *-ül* (in *vizül*) should not be analyzed as employing the long diphthong <eu>.
- *Morpheme Boundary: hűl-* (verb stem) + *-j* (3rd-person singular imperative/jussive marker) → Hűljön (meg) a véred! – May your blood grow cold! *[11]*
- *Sound Change:* The direction of total assimilation is reversed relative to today's standard language: instead of being regressive, it is progressive: /l + j/ → [l:], marked in writing as -ll, e.g., h<ű>l>l.
- *Prosody:* The syllable sequence is heavy + light + light. The <eü> diphthong is not called for here, because it would make the metrical foot one syllable longer. The variant *vizeül*, proposed by Jakubovich–Pais for 'vízként' ('as water'), fails to meet the requirements of dactylic rhythm (*vizül → vizeül*). In stanza 3/b, the rhyme is a suffixal (grammatical) rhyme: *mézül ~ vízül. [6]*

5. *Semantic Nuance*

Ruit and *fluit* in Latin denote rapid flow, not falling. The Son's life-giving blood comes to resemble water. According to late medieval medicine, the heat of outflowing blood is quickly lost. Mary's question— Véred hűljön meg, mint a víz? – Shall your blood grow cold, like water?—voices, in horror, this immediate cooling. By contrast, *hull* signifies mere gravitational falling and conveys neither the surge nor the loss of heat.

6. *Summary for rhythm-based phonology*

Aspect	Rhythm-driven decision	Phonological description
Syllabification:	*hűlj* <hűll>	1 syllable (1 heavy)
Syllable type:	CVCC	full syllable type
Mora pattern:	– ∪ ∪	The word form hűlj provides the 4th dactyl.
Morpheme boundary:	hűl- + -j	/l + j/ → [l:] progressive total assimilation
Meaning:	'hűljön meg', 'veszítse el melegét' – 'may it grow cold', 'lose its heat'	The semantic shift—namely, the cooling of the flowing blood—accords with the *ruit/fluit* dynamism.

IV. LINGUISTIC MELODY

4.1 The Iambic Meter of the Fragment a Hun Trilógia (The Hun Trilogy)—János Arany's Late Epic Diction

Lajos Áprily's 7- and 6-syllable iambic line was not unknown in earlier Hungarian poetry; cf. the twentieth-century Hungarian poet Lajos Áprily's poem titled *Ómagyar Mária-siralom* from 1938. *[12]*

Hul-ló | köny-nyem | mu-tat-|ja ‖ s szün-te-|len só-|ha-jom, a
– – | – – | ∪ – | ∪ ‖ – ∪ | – – | ∪ –
hogy meg-|gyö-tör |se-bé-|vel ‖ a bel-|ső fáj-|da-lom. a
– – | ∪ – | ∪ – | ∪ ‖ ∪ – | – – | ∪ –

He had already used it in the epic *The Hun Trilogy*, where he versified his hero Attila in alexandrines composed of two four-and-a-half-foot iambic lines. One might suppose that Arany took his model from the *Nibelungenlied*, since he translated four stanzas from the *Nibelungenlied*-the Twenty-Fifth Adventure-which is related to PAN stanzas 6a–6b, dividing the closed lines with diaereses *[13]:*

The Twenty-Fifth Adventure

Ott állt | sze-gény |pa-pocs-|ka, ‖ ráz-ván | lucs-kos | me-zét. a
– – | ∪ – | ∪ – | – ‖ – – | – – | ∪ –
Meg-tud-|ta Ha-|gen eb-|ből, ‖ hogy nem | ha-zug | be-széd a
– – | ∪ – | ∪ – | – ‖ – – | ∪ – | ∪ –
A bi-|zo-nyos | ha-lál, | mit ‖ jó-solt | a hab-|le-ány; b
∪ ∪ | ∪ – | ∪ – | – ‖ – – | ∪ – | ∪ –
Gon-dol-|ta: „mind | e hő-|sök ‖ ott vesz-|nek, i-|ga-zán." b
– – | ∪ – | ∪ – | – ‖ – – | ∪ ∪ | ∪ –

PAN 6a–6b in Hungarian

O, í-| gaz Sí-| me-on-| nak ‖ bez-zeg | sza-va | é-re, a
∪ –́ | – –́ | ∪ – | – – ‖ – – | ∪ ∪ | – –
En ér-| zem ez | bú-tő-| röt, ‖ kit né-| ha í-| gé-re. a
∪ – | ∪ – | – – | – ‖ – –́ | ∪ –́ | – –
Si-ral-|mam, fo-|há-sza-|tom ‖ te-rít-|he-tők | kívül, b
∪ – | – ∪ | – ∪ | – ‖ ∪ – | ∪ – | – –
En jon-|hom-nak | bel bú-|ja, ‖ ki söm-|ha nem | hé-vül. b
– – | – – | – – | – ‖ ∪ – | ∪ – | – –

In the *Nibelungenlied* quotation, we see that the iambic anacrusis is absent at the beginning of the line or after the main caesura—something that is very much present in ÓMS and is characteristic of János Arany's ingenious late epic poetry:

From the first part of the Csaba trilogy

A mult | i-dők | ho-má-|lyán ‖ meg-szó-|lal egy | re-ge a
∪ – | ∪ – | ∪ – | – ‖ – – | ∪ – | ∪ –

Mint el-|ha-ló | menny-dör-|gés, ‖ fü-lem-|ben é-|ne-ke; a
– – | ∪ – | – – | – ‖ ∪ – | ∪ – | ∪ –
Mint nagy | vi-zek | mo-raj-|ját, ‖ mely-től | zúg a | va-don, b
– – | ∪ – | ∪ – | – ‖ – – | – ∪ | ∪ –
Vér-rel | fo-lyó | na-pok | bús ‖ pa-nasz-|szát hall-|ga-tom. b
– – | ∪ – | ∪ – | – ‖ ∪ – | – – | ∪ – [13]

The iambic anacruses in János Arany's *The Hun Trilogy* can be observed in *The Csaba Trilogy, Part I* [First Canto: Átilla and Buda; Second Canto: Átilla Goes A-Hunting], as well as in *The Csaba Trilogy, Part III* [First Canto: Átilla Dies (Fragment from the First Canto); Second Canto: The Bride's Awakening; Third Canto: Zoárd's Counsel; Fourth Canto: Átilla on the Pall].

From the above list, it may be concluded that an iambic anacrusis—at the beginning of the line or in the metrical foot following the main caesura—is by no means alien to János Arany; cf. Zoltán Kodály's opposing view of the iambic verse forms of Arany and Petőfi: "with a heavy beginning, completely omitting the iambic cadence." *[14]*

From the dictionary section of *The Reverse Dictionary of Hungarian Word Endings* [15], I syllabified the headwords: *58,301* dictionary entries. In these, I counted the possible metrical feet and found a total of *164,234*. I have summarized the relative proportions of the metrical feet in the table below:

Metrical Feet		
Total (count):	164.234	100%
Pyrrhic	30.792	18,75%
Spondee	39.877	24,28%
Iamb	32.543	19,82%
Trochee	42.750	26,03%
Anapest	7.701	4,69%
Dactyl	10.571	6,44%

From the table, it is evident that trochees have an advantage over iambs; nevertheless, the number of iambic feet is still high—nearly 20% of the 164,234 metrical feet examined. *[16]*

If János Arany had known the text of PAN, he would have been able to translate it in iambs, as the author of ÓMS did. Indeed, I am of the view that Arany may have known the PAN text in fragmentary form-from two seventeenth-century hymnals, the Kájoni *Kancionále* and the *Cantus Catholici*. The complete version of PAN might also have reached him, via the Library of the Hungarian Academy of Sciences, from *Schauspiele des Mittelalters*. *[17]*

János Arany was not only a poet; he also composed melodies for his own poems and for those of others. His original melodies were published by Zoltán Kodály on the basis of István Bartalus's work. *[14]*

Bartalus's manuscript, in 1875, mentions the first piece to be notated—one that is important for us from the perspective of ÓMS. *[18]*

Zoltán Kodály considers this five-line melody unusual in its form. *[14]*

This is confirmed by András L. Kecskés, an expert in early music: "The textless piece's unusual major–minor melody stands out among Arany's compositions, not to mention the possible 4×7 verse form" *[19]*; cf. the 5×7 verse form.

According to Kodály, the melody's text is a well-known snippet of a humorous folk rhyme, e.g.:

Zsidó, zsidó, vak apád !
M'é' nem öszöl szalonnát ?
Lá'd a magyar mögöszi,
Drága pé'zön mögvöszi ;
De a zsidó nem öszi,
Mer' a fene mögöszi ! *[20]*

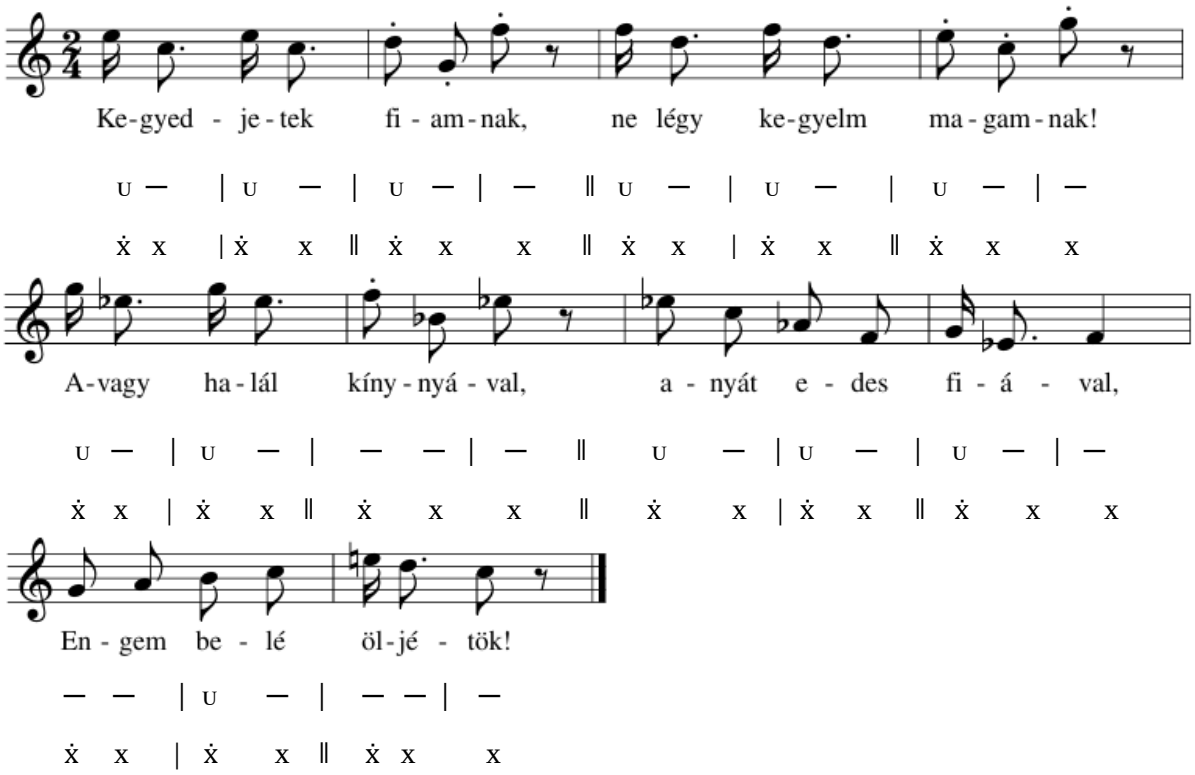
All the same, if the text of PAN surfaced in the Kájoni *Kancionále* and the *Cantus Catholici*, then for János Arany's song composition the Hungarian text of ÓMS 8b could likewise have been recalled in the Latinate 5×7 stanza even in the nineteenth century: *Kegyedjetek fiamnak, | ne légy kegyelm magamnak!* – "Show favour to my son | be not gracious to me!"

cf. the two iambic rhythms, namely in phrases built on the shared root "kegy-":

1. The one in *Bolond Istók*: "kegyelmes herceg" (*gracious prince*)
2. And the one in ÓMS.: "kegyedjetek fiamnak" (*show favour to my son*). *[13]*

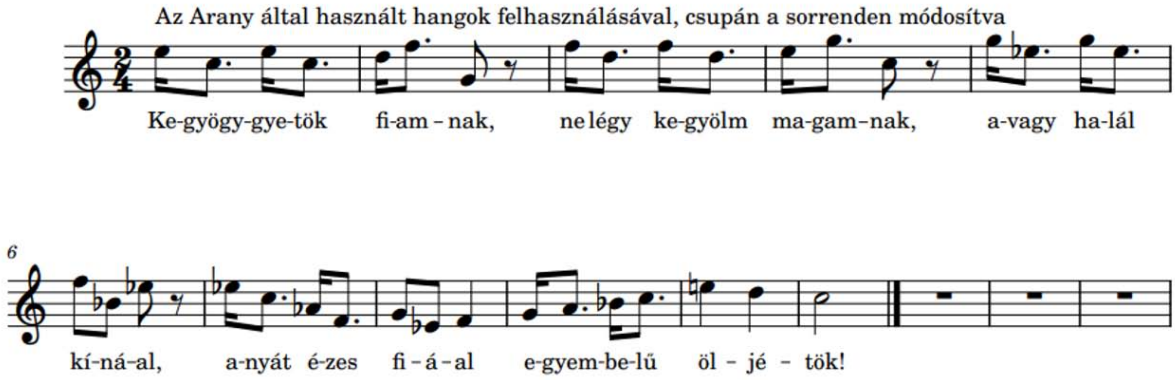
The first line of the original PAN 8a–8b sequence is a reprise (recurring) melody; therefore, the stanza is not five lines but, musically, four- something János Arany did not take into account.

Instead of pairing János Arany's melody with a humorous folk rhyme, I assigned to it the text from ÓMS—the catchy God-lament of the Virgin Mary—from stanza 8b, as in the score edition *[14]*; cf. Bartalus's manuscript II. *[18]*:

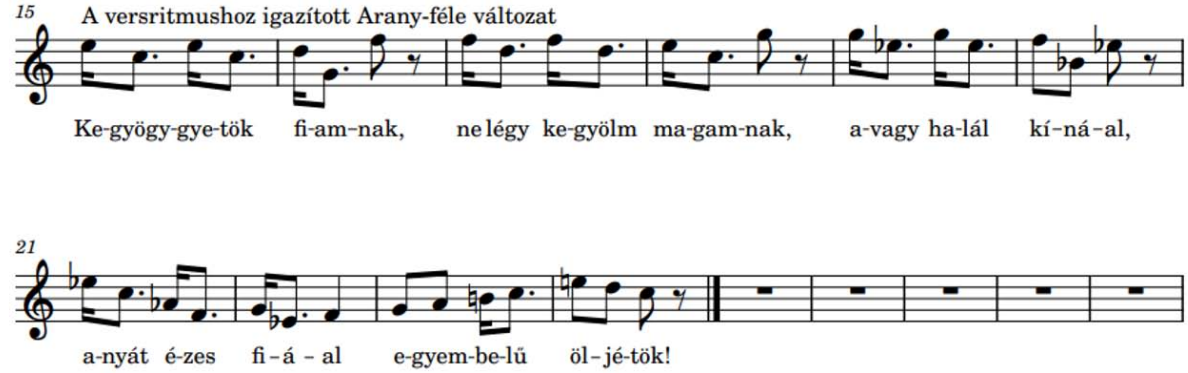


Szilárd Ferenc Kovács, a church musician, adapted Arany's scholarly version to the verse rhythm, then—using the pitches Arany employed—brought the melody closer to the desired character (secular versus ecclesiastical). The score is published with the oral approval of Szilárd Ferenc Kovács.

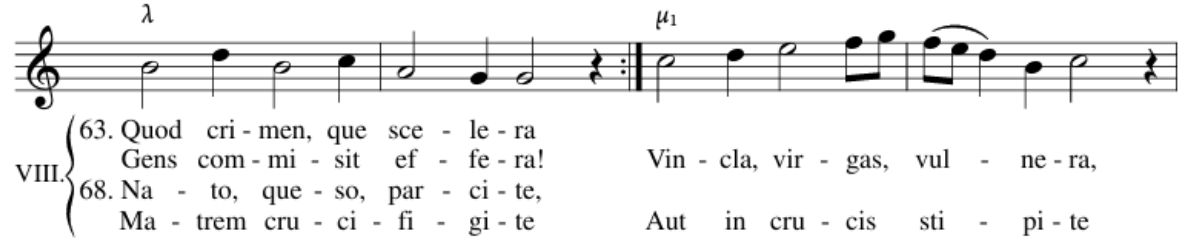
1. Using the pitches employed by János Arany, with only the order modified.



2. The Arany version adjusted to the verse rhythm.



In the PAN 8a–8b sequence, the Latin trochaic cadence and the reprise melodic section are clearly traceable; score edition *[21]*:



1st line of verse:

63. — ∪ | — ∪ | — ∪ | — ‖
ẋ x | ẋ x ‖ ẋ x x ‖

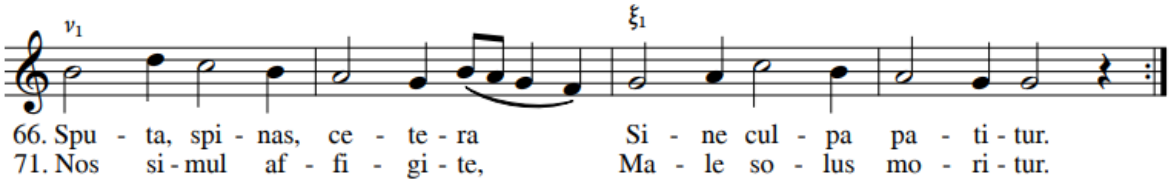
64. — ∪ | — ∪ | — ∪ | — ‖ 65. — ∪ | — ∪ | — ∪ | —
ẋ x | ẋ x ‖ ẋ x x ‖ ẋ x | ẋ x ‖ ẋ x x

2nd line of verse:

68. — ∪ | — ∪ | — ∪ | — ‖
ẋ x | ẋ x ‖ ẋ x x ‖

69. — — | — ∪ | — ∪ | — ‖ 70. — ∪ | — — | — ∪ | —

ẋ x | ẋ x ‖ ẋ x x ‖ ẋ x | ẋ x ‖ ẋ x x



66. Spu - ta, spi - nas, ce - te - ra Si - ne cul - pa pa - ti - tur.
71. Nos si - mul af - fi - gi - te, Ma - le so - lus mo - ri - tur.

1st line of verse:

66. — ∪ | — ∪ | — ∪ | — ‖ 67. — ∪ | — ∪ | — ∪ | —

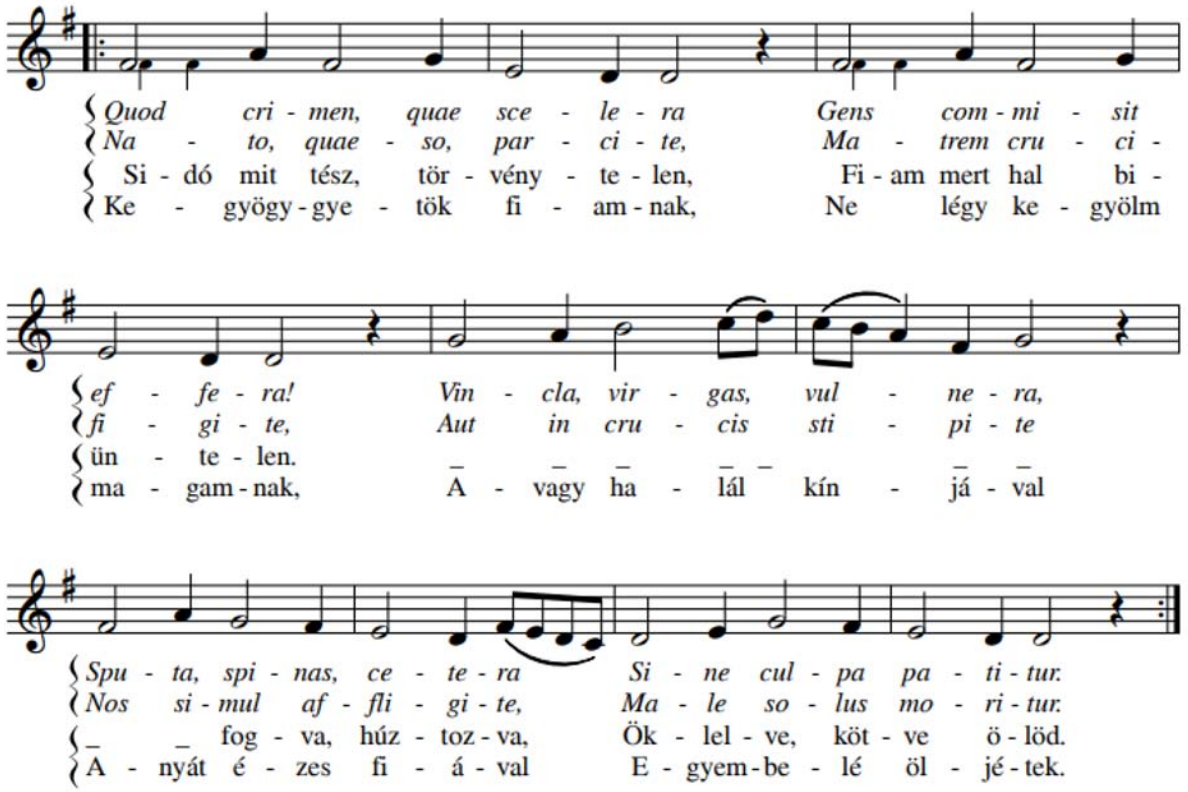
ẋ x | ẋ x ‖ ẋ x x ‖ ẋ x | ẋ x ‖ ẋ x x

2nd line of verse:

71. — ∪ | — — | — ∪ | — ‖ 72. — ∪ | — — | — ∪ | —

ẋ x | ẋ x ‖ ẋ x x ‖ ẋ x | ẋ x ‖ ẋ x x

János Arany's composition—preserved without text but connectable to the iambic rhythm of ÓMS—supports the hypothesis that Bence Szabolcsi rhythmized the principal line of ÓMS incorrectly from a musical standpoint ("*Kegyöggyetök fiamnak, Ne légy kegyölm magamnak*" – "Show favour to my son | be not gracious to me!"), since it is not sufficient to consider PAN alone, score edition *[22]:*

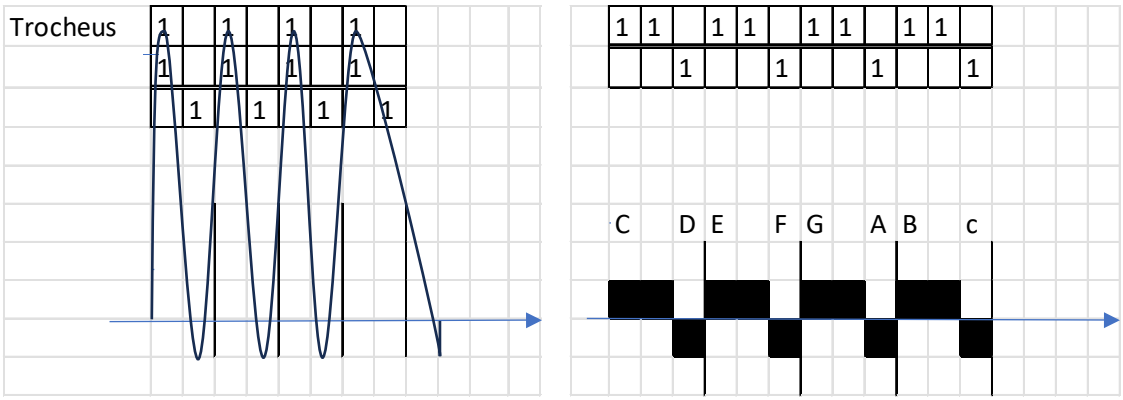


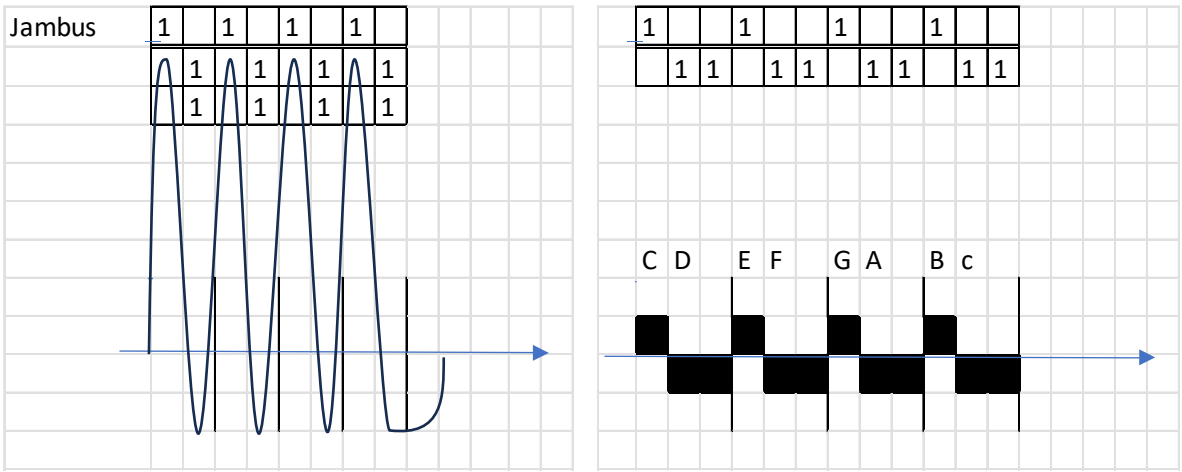
4.2 Representing Meter

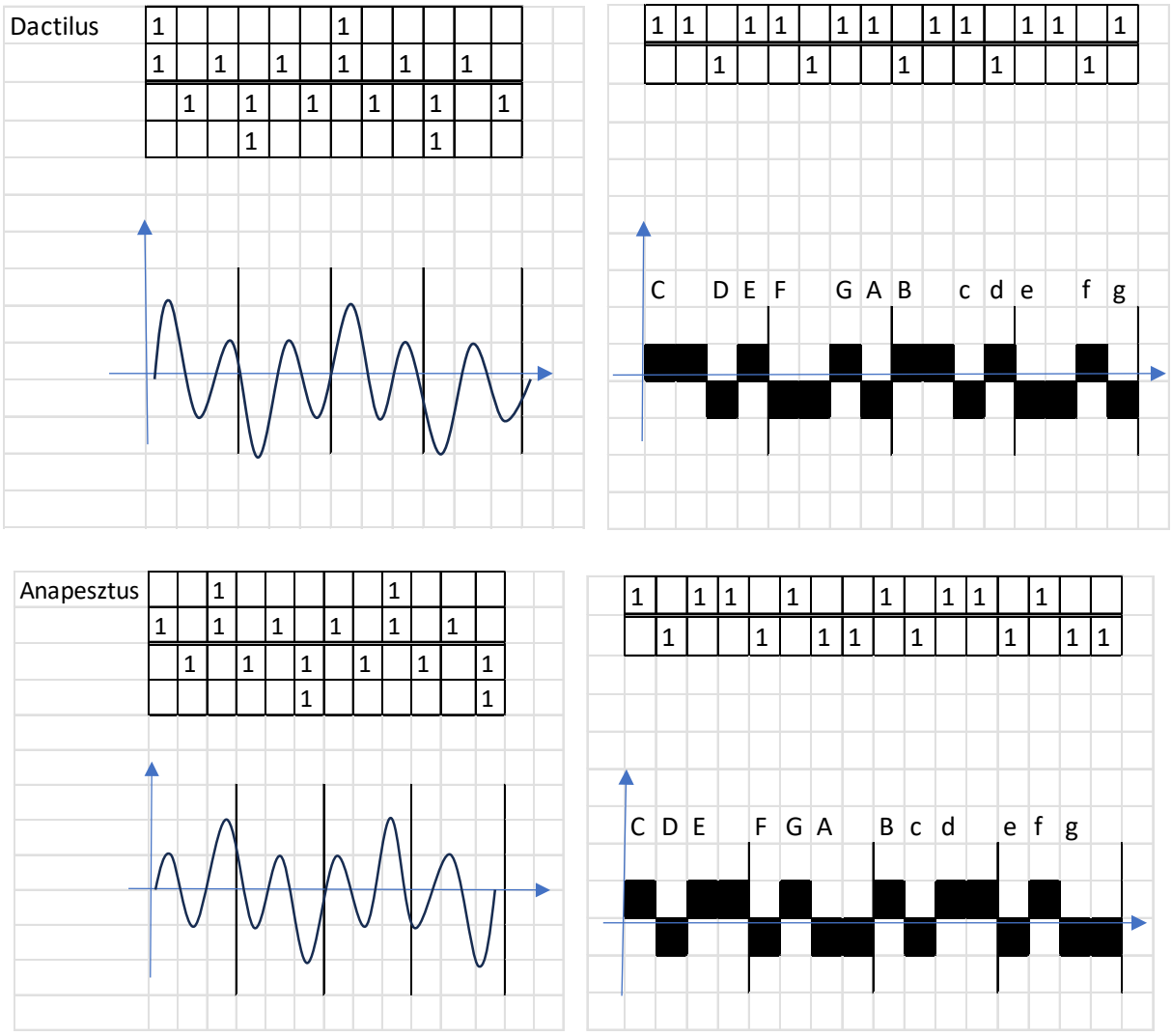
I represent a meter in an abstract manner. It is well known that in quantitative versification the duration value of a short syllable (mora) is one unit (υ = 1), while the duration of a long syllable is two units (1+1).

a) In the left-hand plots below, the horizontal axis shows the number of syllables, and the vertical axis shows the pronunciation time of each syllable. Vertical lines serve to delimit the quantitative metrical feet. Thus, the foundations of metrics do not in themselves address the relationship between the verse line and the melody. *[23]*

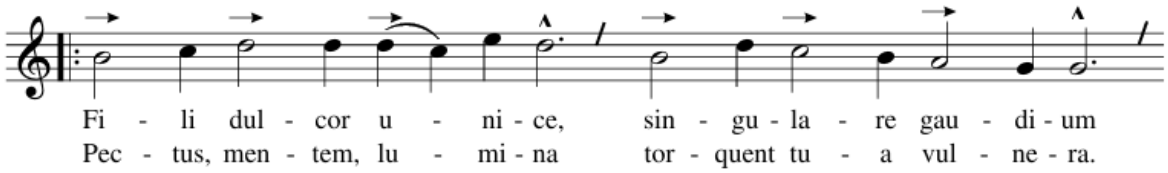
b) In the right-hand plots below, the horizontal axis shows quantitative syllabification (long–short, short–long), while the vertical axis shows pitch (melody), labelled by note name. In this way, the theory substantiates the relationship between language and melody.







A characteristic property of a pitch series is that, under certain conditions, rhythm emerges—on the basis of temporal relations and stress hierarchies. The example below illustrates, in the PAN 2a–2b sequence, the four-and-a-half-foot trochaic meter and the three-beat heptasyllabic rhythm, as well as the melody with long and short syllables; score edition. *[24]*



1st line of verse:

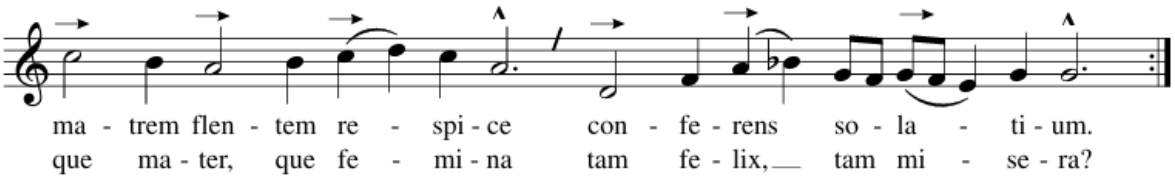
— ∪ | — ∪ | — ∪ | — ‖ — ∪ | — ∪ | — ∪ | —

ẋ x | ẋ x ‖ ẋ x x ‖ ẋ x | ẋ x ‖ ẋ x x

2nd line of verse:

— — | — — | — ∪ | — ‖ — — | — ∪ | — ∪ | —

ẋ x | ẋ x ‖ ẋ x x ‖ ẋ x | ẋ x ‖ ẋ x x



1st line of verse:

— — | — — | — U | — ‖ — U | — U | — U | —

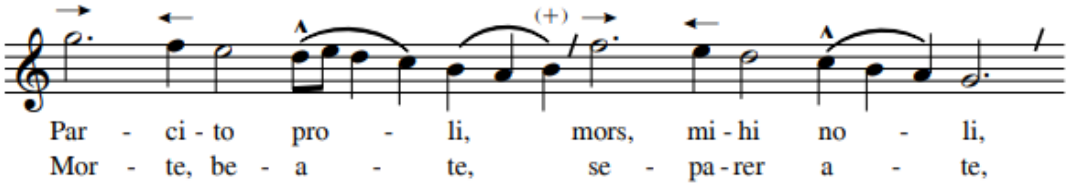
ẋ x | ẋ x ‖ ẋ x x ‖ ẋ x | ẋ x ‖ ẋ x x

2nd line of verse:

— — | — U | — U | — ‖ — — | — — | — U | —

x ẋ | x x | ẋ x x ‖ x ẋ | x x | ẋ x x

In the 7a–7b sequence, Jammers enforces the rhythm of the Adonic lines, with melodies:



1st line of verse:

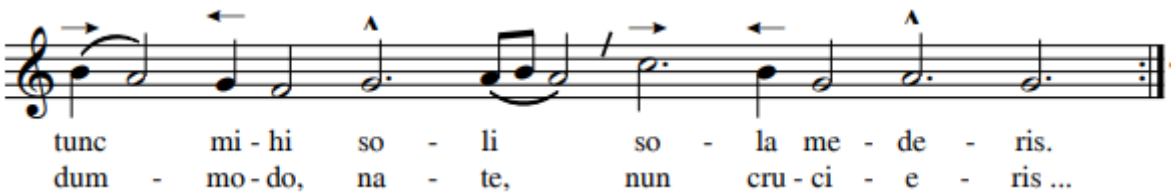
— U U | — U ‖ — U U | — U

ẋ x x ‖ ẋ x ‖ ẋ x x ‖ ẋ x

2nd line of verse:

— U U | — U ‖ — U U | — U

ẋ x x ‖ ẋ x ‖ ẋ x x ‖ ẋ x



1st line of verse:

— U U | — U ‖ — U U | — U

ẋ x x ‖ ẋ x ‖ ẋ x x ‖ ẋ x

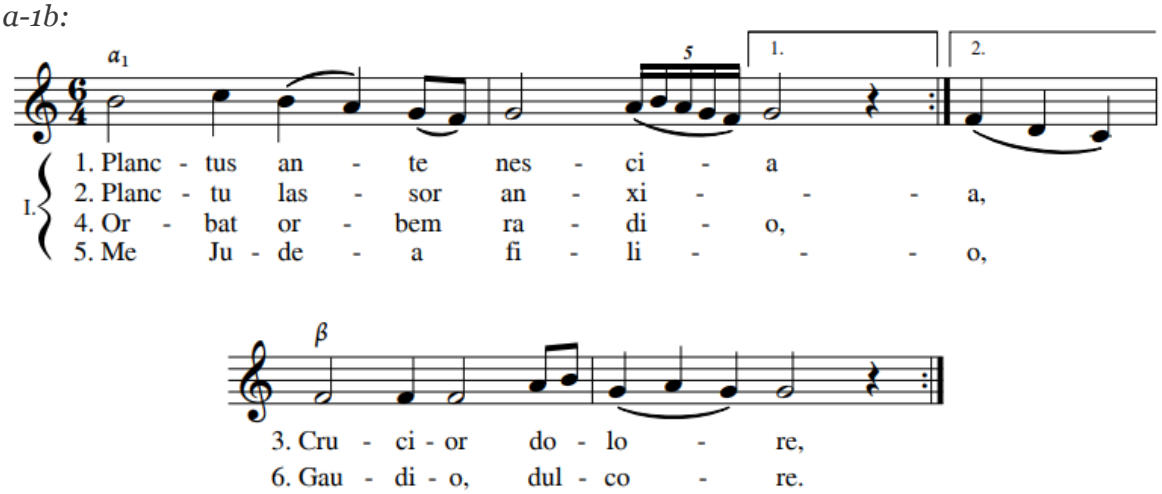
2nd line of verse:

— U U | — U ‖ — U U | — U

ẋ x x ‖ ẋ x ‖ ẋ x x ‖ ẋ x

Using the melodies of PAN 1a–1b and 6a–6b, as well as 3a–3b, I demonstrate the opposition between the prevailing trochaic and dactylic lines—an opposition that ÓMS also preserved, score edition. *[21]*

PAN 1a-1b:



Lines 1–2 of the verse:

— U | — U | — U | — ‖
ẋ x | ẋ x ‖ ẋ x x ‖
— U | — U | — U | —
ẋ x | ẋ x ‖ ẋ x x

3rd line of the verse:

— U | — U | — —
ẋ x | ẋ x ‖ ẋ x

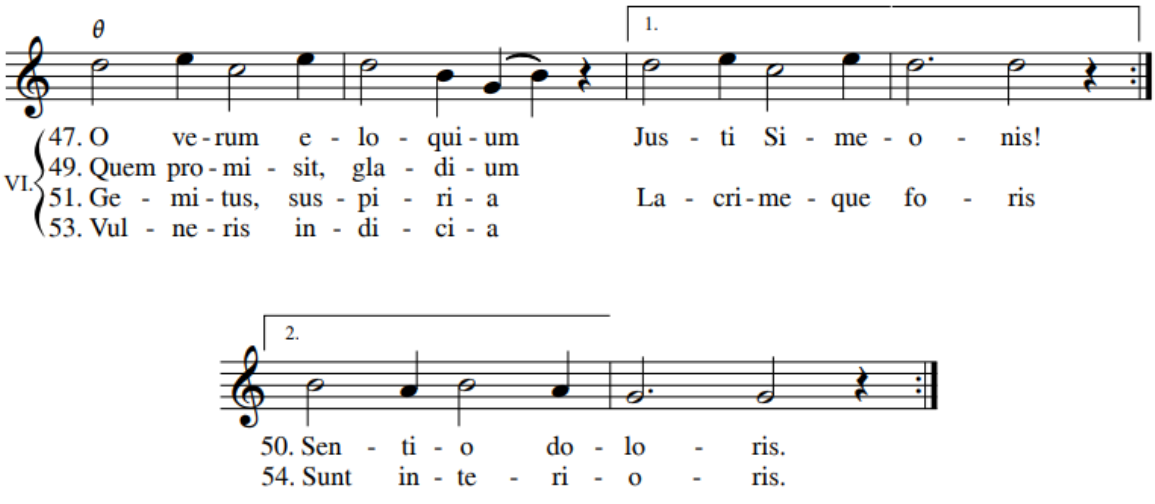
Lines 4–5 of the verse:

— U | — U | — U | — ‖
ẋ x | ẋ x ‖ ẋ x x ‖
— U | — U | — U | —
ẋ x | ẋ x ‖ ẋ x x

6th line of the verse:

— U | — U | — —
ẋ x | ẋ x ‖ ẋ x

PAN 6a-6b:



Lines 47–48 of the verse:

— ∪ | — ∪ | — ∪ | — ‖ — ∪ | — ∪ | — — |

ẋ x | ẋ x ‖ ẋ x x ‖ ẋ x | ẋ x ‖ ẋ x

Lines 49–50 of the verse:

— ∪ | — – | — ∪ | — ‖ — ∪ | — ∪ | — — |

ẋ x | ẋ x ‖ ẋ x x ‖ ẋ x | ẋ x ‖ ẋ x

Lines 51–52 of the verse:

— ∪ | — – | — ∪ | — ‖ — ∪ | — ∪ | — — |

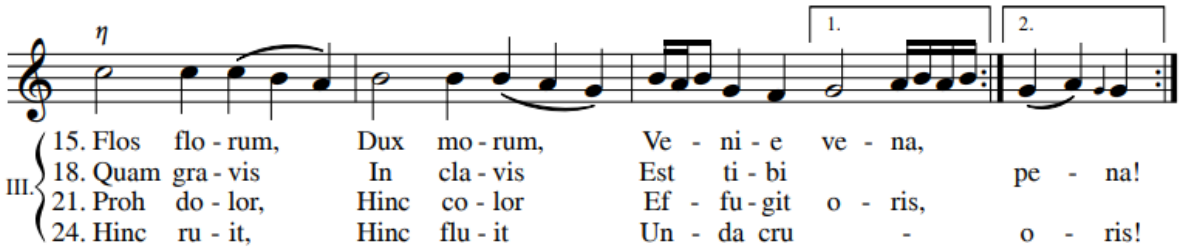
ẋ x | ẋ x ‖ ẋ x x ‖ ẋ x | ẋ x ‖ ẋ x

Lines 53–54 of the verse:

— ∪ | — – | — ∪ | — ‖ — – | — ∪ | — — |

ẋ x | ẋ x ‖ ẋ x x ‖ ẋ x | ẋ x ‖ ẋ x

PAN 3a-3b:



Lines 15–20 of the verse:

— ∪ ∪ ‖ — ∪ ∪ ‖ — ∪ ∪ | — ∪

— ∪ ∪ ‖ — ∪ ∪ ‖ — ∪ ∪ | — ∪

ẋ x x ‖ ẋ x x ‖ ẋ x x ‖ ẋ x

ẋ x x ‖ ẋ x x ‖ ẋ x x ‖ ẋ x

Lines 21–26 of the verse:

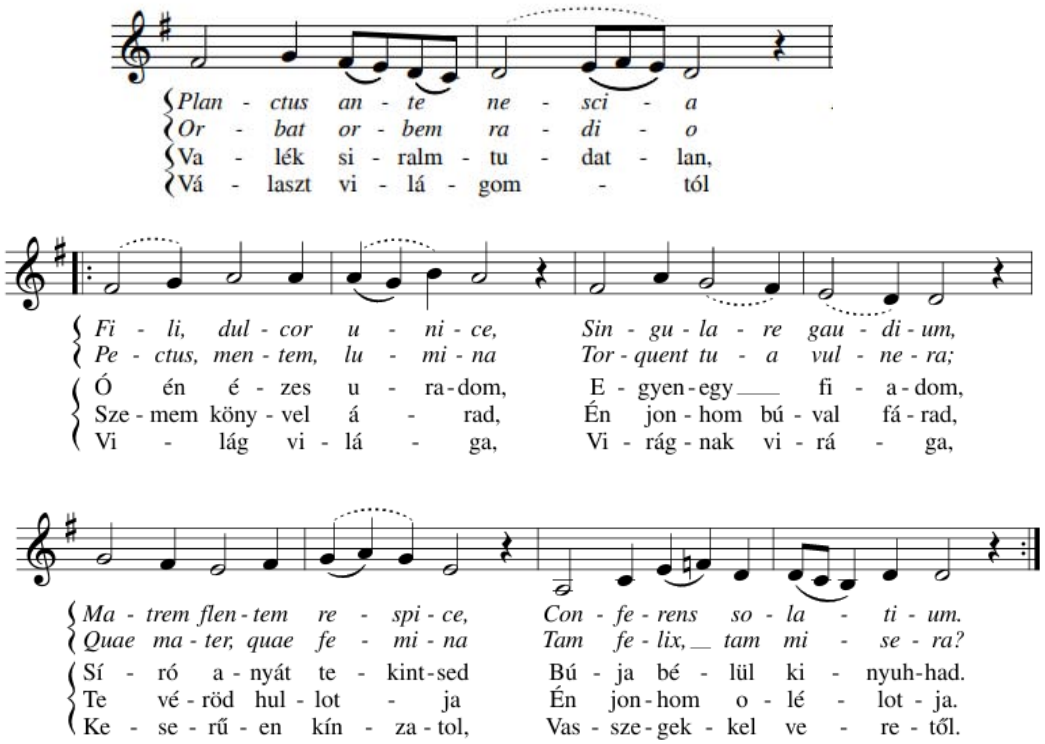
— ∪ ∪ ‖ — ∪ ∪ ‖ — ∪ ∪ | — ∪

— ∪ ∪ ‖ — ∪ ∪ ‖ — ∪ ∪ | — ∪

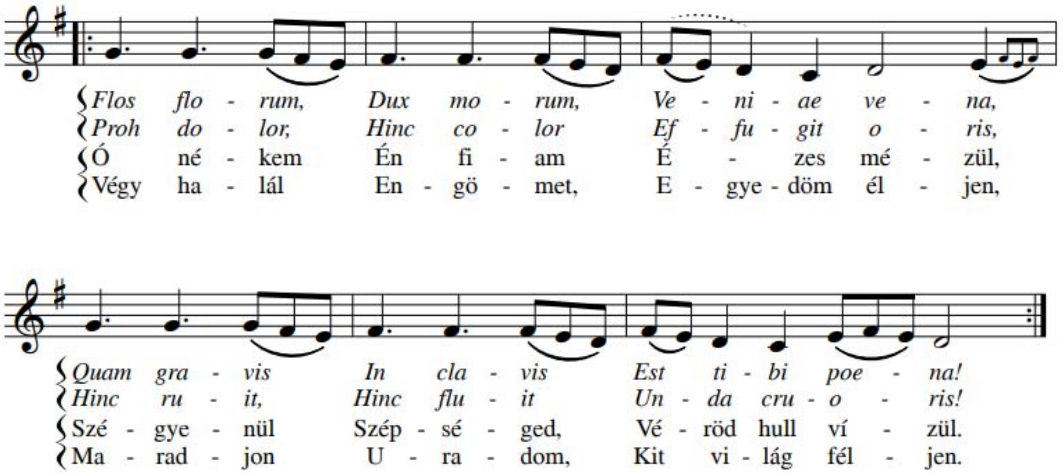
ẋ x x ‖ ẋ x x ‖ ẋ x x ‖ ẋ x

ẋ x x ‖ ẋ x x ‖ ẋ x x ‖ ẋ x

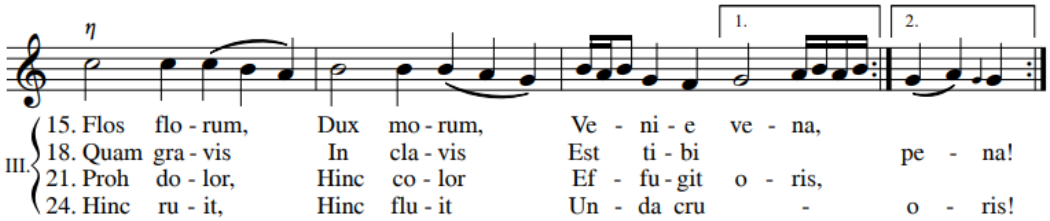
Under the influence of Bence Szabolcsi, to this day, when ÓMS is performed in song, a trochee is incorrectly imposed on an iambic foot; score edition [22]:



In the short lines, with respect to the two-layer versification of PAN and ÓMS, Bence Szabolcsi fails to recognize the dactyl; score edition **[22]**:



cf. **[21]**:



The flagship of ÓMS is music. In the company of sacred folk hymns, the largest-scale musical work known today about ÓMS was composed by Ferenc Ottó, a student of Zoltán Kodály. His 80-minute oratorio *The Hungarian Passion* (formerly *Mary's Lament*) is forthcoming—cf. *[25]*

V. SUMMARY

a. Thesis And Interpretive Framework

The study's opening sentence—"The relationship between consciousness and linguistic data to formalize, using symmetry principles I have identified to formalize that relationship"—*is interpreted to mean* that consciousness as a biological fact and formalized linguistic data are mirror images of one another: the organizing principles of conscious processing appear in linguistic structures as systems of symmetry, and conversely, the symmetries demonstrable in language reflect the organization of consciousness. The study makes this mirror-like correspondence accessible and measurable through a symmetry-based formalization. The symmetry-theoretical framework rests on three coequal pillars: *(1) machine encoding* (based on syllable and word position, using a center–periphery model), *(2) rhythm-based phonology* (patterns of quantitative meter and stress in historical texts), and *(3) linguistic melody* (the alignment of intonation and meter). The integration of these three perspectives provides the internal logic for formalizing the "consciousness–language" relationship.

b. Theoretical Novelty and Model

Symmetry as a central concept is not merely a descriptive metaphor: explicit encoding schemes and positional labels carry the structure through at the database and algorithmic levels. The core *center–periphery model* represents the vowel (V) as the center and the consonants (C) as concentric, symmetrical layers (C<C<C<V>C>C), which brings both the syllable types and the within-syllable positions of phonemes into a unified notation.

The study also introduces a formal, "search-styled" relational notation:

CONSCIOUSNESS (search) = [1 : 2] : [3 : 4], where (1) is the syllable template, (2) the syllable's word-initial/medial/final status, (3) the relative position within the word, and (4) the word category by syllable count. This compact notation expresses the mutual mapability of linguistic rhythm and structural placement—*a mirror correspondence between the "biological" organization of consciousness and the "formal" linguistic data.*

c. Methodology: Three Coequal Pillars

d. Machine encoding — symmetry-based classification

The classification of linguistic units into word types by syllable count (D, A, B1…B8) and the resulting set of 55 derived syllable types is designed to make rhythmic and positional information algorithmically manageable. Syllables fall into five main structural classes (reduced, inorganic, open, closed, full), with a detailed repertory of CV patterns, all unified by the center–periphery model. The positional encoding of phonemes (e.g., CVC-1, CVC-2; CVCC-1/2/3, etc.) and the systematic labelling of syllable types (CVCC:D; CVC:A1; …) capture both internal composition and function within the word; this dual code produces a queryable, scalable database for speech-to-text systems and computational linguistics.

1. Rhythm-based phonology-historical case studies

The versification of the *Ómagyar Mária-siralom* (ÓMS) is compared on the basis of the critical edition of the Latin *Planctus ante nescia* (PAN): while the dominance of Latin trochaic formations is well known, the Hungarian text exhibits iambic solutions, and dactyl–dactyl correspondences can also be demonstrated. Two detailed philological–phonological case studies show how meter decides between graphic variants: (1) in the case of *biüntlen/bűntelen,* iambic fit supports the segmentation *bi-ünt^{e}-len;* (2) in the case of *hűlj/hull,* both progressive /l + j/ assimilation and the dactylic pattern are consistent with the reconstruction *hűlj,* while *"hull"* does not violate the rhythm either but carries a different semantic drift. These examples demonstrate the role of rhythm as arbiter amid orthographic uncertainties and show the practical disambiguating power of formal metrical description.

1. *Linguistic melody — aligning meter and intonation*

   The study also provides an abstract, functional representation of meter (duration–syllable functions, pitch–time mappings), and through historical examples—including János Arany's late epic diction and melodic materials-demonstrates how pitch contours align with metrical feet and caesuras. In certain PAN sequences, the separately discussed trochaic and dactylic lines, as well as the melodic linkages of Adonic formulas, shed light on a two-level organization (metrical and melodic); this duality further justifies a reconfiguration of ÓMS performance practice (e.g., a critique of traditional trochaic "overlays" imposed on iambic principal lines).

IV. RESULTS AND CONTRIBUTIONS

- *Unified Encoding Architecture:* The study develops a compact, formal notation system that can handle, within a single framework, the internal structure of syllables, positions within the word, and rhythmic aspects. This notation directly "maps" the structures of conscious linguistic composition onto the raw data—and, conversely, makes it possible to read back the organizing principles of consciousness from the formalized data structures (a mirror-like correspondence).
- *Rhythmic disambiguation and reconstruction:* Meter-driven analysis makes it possible to resolve doubtful forms (e.g., *biüntlen / bűntelen*; *hűlj / hull*); in phonological–morphological boundary cases, rhythm signals the preferred reading. This procedure is directly applicable in philological reconstructions and critical editions.
- *Modelling the Rhythm–Melody Interface:* The formal linkage of duration and pitch makes prosodic patterns quantifiable. Examples demonstrating the interaction between linguistic melody and meter (e.g., PAN 1a–1b, 6a–6b, 3a–3b; Adonic formulas) also cast the performance rhythm of ÓMS in a new light.
- *Empirical Background and Typological Balance:* The metrical-foot distribution report (a foot count in the tens of thousands over dictionary data) statistically confirms that the iamb is not a marginal phenomenon (≈20%) within the Hungarian rhythmic inventory, even though the trochee's advantage is demonstrable; this aligns with the picture of rhythmic diversity found in Hungarian historical material.

VI. PRACTICAL APPLICABILITY

The formalized, symmetry-based encoding offers added value across multiple application areas: (1) speech recognition—by incorporating rhythmic and positional features, more robust acoustic–language models can be built; (2) *speech synthesis*—explicit handling of the melody–duration interface yields more natural prosody; (3) *NLP tasks*—automatic detection of morpheme boundaries and syllabification, rhythm-sensitive text generation; (4) *philology and performance practice*—reconstruction of rhythm and melody in historical texts, and correction of misleading traditions (e.g., trochaic overlays). The study emphasizes that seeing the three pillars together is indispensable for designing the linguistic layer of high-speed, scalable "terminal-to-terminal" communication as well (speech-to-text systems, database structures).

VII. CONCLUSION

The results presented argue that the consciousness–language relationship can be formalized along symmetry principles: the rhythmic–structural organization of consciousness and the formal patterns of linguistic data mutually reflect one another. Machine encoding (positional and structural labelling), rhythm-based phonological decision-making, and the model of linguistic melody together offer an integrated, bidirectional framework in which the architecture of conscious language use and the isomorphic (mirror-image) correspondence of formalized linguistic data can be demonstrated and reutilized—both for theoretical understanding and for technological implementations.

The author declares no conflicts of interest.

The analysis above was created with the assistance of ChatGPT.

REFERENCES

1. Halász, I. (2017). Szimmetria, nyelv, irodalom. Záródolgozat. Rátz László vándorgyűlés. 1-12.
2. Csehy, Z. (2014). Virágnak virága. Az Ómagyar Mária-siralom mint ihletforrás a modern és a kortárs költészetben. Magyar Nyelv 120. 2024: 31.
3. Szaszák, Gy. (2009). A prozódia felhasználási lehetőségei a gépi beszédfelismerésben. 1-15.
4. Gragger, Róbert (1923). Ómagyar Máriasiralom. Magyar Nyelvőr 1–13.
5. Gragger, Róbert (1923). Eine altungarische Marien-klage. Ungarische Jahrbücher III. Berlin, 18.
6. Mező, T. (2021). Ómagyar Mária-siralom szövegkönyve. Novum Publishing Kiadó, Neckenmarkt. 104-130.
7. A. Molnár, F. (2005). A legkorábbi magyar szövegemlékek. Olvasat, értelmezés, magyarázatok, frazeológia. Debrecen: (ΑΓΑΘΑ XVII – Nyelvi és Művelődéstörténeti Adattár. Kiadványok 8).
8. Horváth, I. (2015). Ómagyar szövegemlékek mint textológiai tárgyak. Országos Széchényi Könyvtár, Budapest.
9. Mező T. szerk. (2025). Az Ómagyar Mária-siralom megközelítésének új szempontjai. Anthology. Interkulturális Kutatások Ltd., Hajdúböszörmény.
10. Kőszeghy, P.-Szabó, G. (1986). Gyarmati Balassi Bálint Énekei. Szépirodalmi Kiadó, Budapest
11. Mészöly, G. (1956). Ómagyar szövegek nyelvtörténeti magyarázatokkal. Tankönyv-kiadó, Budapest.
12. Áprily, L. (2006). Összes költeményei. Osiris Kiadó, Budapest. 383-4.
13. Arany, J. (2006). Összes költeményei I-II. Compiled and annotated by Márton Szilágyi. Osiris Kiadó, Budapest. 583-4., 636., 907.
14. Kodály, Z.-Gyulai, Á. (1952). Arany János népdalgyűjteménye. Akadémiai Kiadó, Budapest. 181-2., 183., 190.
15. Papp, F. (1969), (1994). A magyar nyelv szóvégmutató szótára. Akadémiai Kiadó, Budapest. 49-537.
16. Mező, Tibor (2014). A magyar nyelv szótagtára. Manuscript, Pomáz. 175.
17. Mone, F. J. (1846), (1852). Schauspiele des Mittelalters. Verlag von J. Vensheimer, Mannheim. 362-6.
18. Bartalus, I. (1875). Bartalus kézirata I-II. Budapest. 427.
19. L. Kecskés, A. (2018). A gitáros muzsikus. Szentendre. 88.
20. Kálmány L. (1881). Szeged népe III. 176. vers, 113.
21. Gennrich, F. (1932). Grundriß einer Formenlehre des mittelalterlichen Liedes als Grundlage einer musikalischen Formenlehre des Liedes. Max Niemeyer Verlag, Halle (Saale). 145., 143-5., 144.
22. Szabolcsi, B. (1947). A magyar zenetörténet kézikönyve. Magyar Kórus, Budapest, 1947. 10*
23. László, Zs. (1985). Költészet és zeneiség. Prozódiai tanulmányok. Akadémiai Kiadó, Budapest. 23.
24. Jammers, E. (1975). Aufzeichnungweisen der einstimmigen außerliturgischen Musik des Mittelalters. Arno Volk-Verlag – Hans Gerig KG, Köln. 4.89-4.90.
25. Ottó, F. (1942). Vallomás a Mária-Siralomról. Vigilia. No. 8. 314-9.