GÁBOR PALKÓ

THE RELATIVE FREQUENCY OF INNER-LIFE VERBS AS SIGNIFIER OF CHANGE IN 19TH- AND 20TH-CENTURY FICTION? DISTANT READING OF A CORPUS OF HUNGARIAN NOVELS[1]

Introduction

In this study, I aim to integrate the investigation of historical developments and specific characteristics of prose fiction from the past two centuries of Hungarian literary history, with a computational linguistic and statistical analysis of a corpus comprising Hungarian novels. Additionally, I situate this methodological integration within the broader landscape of scholarship pursuing similar objectives. Before outlining the methodology and presenting the results of my research, I provide a concise theoretical reflection on the relationship between digital and analogue literary analysis, which constitutes the theoretical underpinning of this study. Subsequently, I survey the relevant antecedents within Hungarian scholarship, alongside significant international examples and parallels. This overview leads into a detailed account of the corpus development and analytical methods, concluding with a discussion of the study's key findings.

Literary Studies and Distant Reading: Actual or Potential Influence?

> I am not saying that there are no results at all; but, in general, we must admit that they [computational literary studies] have a limited impact in the mainstream literary-critical or historical discourse (especially if we compare the momentum of other approaches in the past, say Structuralism, Deconstruction, Cultural studies or New Historicism)[2].

Fabio Ciotti's argument is difficult to dispute: there is a significant gap between computational literary studies (CLS) and traditional literary scholarship. Researchers engaged in CLS and scholars employing conventional methodologies strive to satisfy fundamentally different scholarly norms. These differences are notably visible in the editorial policies of scholarly journals. Journals dedicated to

[1] The research was supported by the National Laboratory for Digital Heritage (Project no. 2022-2.1.1-NL-2022-00009). The project has been implemented with the support provided by the Ministry of Culture and Innovation of Hungary from the National Research, Development and Innovation Fund, financed under the 2022-2.1.1-NL funding scheme.

[2] Fabio Ciotti, "Distant Reading in Literary Studies: A Methodology in Quest of Theory", *Testo e Senso*, 2021, 23, pp. 195-213.

DACOROMANIA LITTERARIA, XII, 2025, pp. 155–178 | DOI: 10.33993/drl.2025.12.155.178

digital humanities and CLS (such as the Digital Scholarship in the Humanities or the International Journal of Digital Humanities or, more specifically, the Journal of Computational Literary Studies), typically discourage arguments built exclusively on the user-level application of ready-made tools that operate as "black boxes", where processes remain opaque to users (as is the case with Voyant Tools or Gephi). These journals increasingly demand open publication of data and code and emphasize technological transparency along with critical reflection on digital methodologies. This shift aligns with broader trends towards data-driven scholarship, exemplified by the institutional pressure to comply with FAIR principles, exerted through research funding policies across nearly all academic fields.

In literary studies specifically, an additional factor warrants attention regarding the relationship between computational methods and traditional research approaches. The "digital turn" in literary studies, predicted and celebrated by some, yet deemed non-existent or incomplete by others, coincides with a crisis of legitimacy most accurately described by the concept of "reproducibility".

In 2019 when Nan Z. Da's high-profile critique attempted to reproduce prominent computational literary studies and found many results could not be duplicated, calls intensified for greater methodological rigor and transparency. Reproducibility has since become both a methodological gold standard and a point of contention – hailed as essential for rigor and scientific accountability, yet seen by some as potentially misaligned with the interpretive ethos and broader humanistic values of literary scholarship[3].

Undoubtedly, the increasing emphasis on the transparency of data and the processes applied to them, as well as the reproducibility of digital methodologies, represents a positive development. This trend aligns closely with methodological standards from hard(er)-science fields closely associated with digital humanities, such as quantitative social sciences and computer science. Nevertheless, the shift toward data awareness in computational literary studies (CLS) can also be interpreted as symptomatic of the aforementioned crisis of legitimacy.

Yet it remains far from certain that this data-driven turn in CLS will bridge the divide between mainstream literary-critical or historical discourse, and the increasingly technical discourse of distant reading. Indeed, the gap appears to be widening for two primary reasons. First, as a consequence of the legitimacy crisis outlined above, the notion of reproducibility inherently privileges methodologies rooted in data-driven scientific thinking. Second, the technologies shaping contemporary digital scholarship, which increasingly dominate the technical toolkit of digital humanities and literary research, are undergoing rapid and transformative

[3] Nan Z. Da, "The Computational Case against Computational Literary Studies", *Critical Inquiry*, 45, 2019, 3, pp. 601-639. From a broader perspective see Thorsten Ries, van Dalen-Oskam, Fabian Offert, "Reproducibility and Explainability in Digital Humanities", *International Journal of Digital Humanities*, 2024, 6, pp. 1-7.

change. Consequently, these technologies – e.g. those enumerated by Ciotti[4] – become increasingly opaque and less accessible to traditional humanities scholars.

Yet there is a third factor contributing to the widening of this divide. Moretti's provocative prediction regarding the global transformation of literary studies as a consequence of distant reading (here understood primarily as quantitative rather than exclusively computational) seems particularly relevant[5]. This logic seems to manifest itself concretely in the current trend within digital humanities, which is characterized by technologically sophisticated arguments that confidently make claims about cultural phenomena – such as literary periods, stylistic movements, or individual authors – about which the scholars themselves lack disciplinary knowledge. Indeed, practitioners of computational reading methods sometimes produce interpretations of texts they cannot read in the original languages. Within traditional literary scholarship, this represents a fundamental breach of scholarly norms. Expertise in traditional literary studies is defined precisely by proficiency in the linguistic and discursive worlds of a given literary period, and, equally importantly, thorough familiarity with the relevant scholarly literature. According to the paradigm of "close reading", scholarly legitimacy is established through publications demonstrating significant expertise in both primary texts and secondary literature.

Equally problematic, and perhaps even irritating, for literary scholars who consider themselves expert interpreters of literary texts, is the use of literary works stripped of their aesthetic function and employed merely as illustrative examples for technologies or methodologies largely disconnected from what Niklas Luhmann describes as the social system of art – the medium through which literary works gain meaning[6].

This study does not aim to offer further explication of the above issues, much less propose a definitive bridging of the gap between computational literary studies (CLS) and traditional literary scholarship. Nevertheless, it seeks to adhere methodologically and argumentatively to the expectations of both domains, reflecting critically on their contrasting norms within the context of the research presented.

The relevance of this endeavour in the context of the current study is heightened by recent developments surrounding openly accessible Hungarian

[4] Ciotti, "Distant Reading", p. 195: "… probabilistic topic modeling, support vector machines, naïve Bayes classifiers, word embedding, and machine learning".

[5] Franco Moretti, "Conjectures on World Literature", *New Left Review*, 2000, 1, p. 57: "Literary history will quickly become very different from what it is now: it will become 'second hand': a patchwork of other people's research, without a single direct textual reading. Still ambitious, and actually even more so than before (world literature!); but the ambition is now directly proportional to the distance from the text: the more ambitious the project, the greater must the distance be".

[6] Niklas Luhmann, *Art as a Social System*. Transl. by Eva M. Knodt, Stanford, Stanford University Press, 2000.

literary corpora from Eötvös Loránd University (ELTE)[7], which have increasingly served as resources for computer-assisted literary analyses. These corpora have also been the basis for studies conducted by scholars unfamiliar with Hungarian literary discourse, by researchers who do not read Hungarian. From the perspective of CLS standards, this is an accepted practice, given that the primary scholarly focus of such research typically lies beyond the aesthetic qualities or linguistic construction of the texts themselves[8]. Conversely, for traditional literary scholars, this methodological practice may represent a significant breach of disciplinary norms.

Previous Research in Hungary

After establishing the theoretical context, it is necessary to examine the historical antecedents of the present research. Just as scholarly handbooks often cite Roberto Busa's Index Thomisticus as one of the earliest significant examples of computer-assisted textual analysis for humanities purposes, the pioneering figure in the computational processing of Hungarian literature is undoubtedly Iván Horváth. His database-building project, initiated in the 1970s, marks a foundational moment in this domain[9].

The historical background of this project is closely related to the sudden emergence of quantitative literary studies at the end of the 1960s. As András Kappanyos aptly observes in a brief essay in 2012[10], the prominence of quantitative methodologies at that time could be interpreted as promising a form of ideological independence from the Marxist orthodoxy imposed by the communist regime. What is particularly interesting, however, is that Kappanyos not only classifies the developments in quantitative literary studies of the 1960s and 1970s as ephemeral

[7] Péter Horváth et al., "ELTE Verskorpusz: a magyar kanonikus költészet gépileg annotált adatbázisa" ["ELTE Poetry Corpus: A Machine Annotated Database of Canonical Hungarian Poetry"], in Gábor Berend et al. (eds.), *XVIII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2022)* [*18th Hungarian Conference on Computational Linguistics*], Szeged, JATEPress, 2022, pp. 375-388; Botond Szemes et al., "Az ELTE Drámakorpuszának létrehozása és lehetőségei" ["The Creation and Potential Applications of ELTE Drama Corpus"], in József Tick, Károly Kokas, András Holl (eds.), *Valós térben, Az online térért: Networkshop 31: országos konferencia* [*In Real Space, for Online Space: Networkshop 31 – National Conference*], Budapest, HUNGARNET Egyesület, pp. 170-178.

[8] See, for example, Christof Schöch, Julia Dudar, Evgeniia Fileva, Artjoms Šeļa, "Multilingual Stylometry: The Influence of Language on the Performance of Authorship Attribution using Corpora from the European Literary Text Collection (ELTeC)", in Wouter Haverals, Marijn Koolen, Laure Thompson (eds.), *Proceedings of the Computational Humanities Research Conference 2024*, Aachen, PublisherCEUR-WS.org, 2024, pp. 386-408.

[9] Zsófia Fellegi, *A digitális filológia Magyarországon: elvek és gyakorlatok* [*Digital Philology in Hungary: Principles and Practices*], Doctoral Thesis, Pázmány Péter Catholic University, Budapest, 2024, p. 17, https://disszertacio.ppke.hu/id/eprint/583/. Accessed March 12, 2025.

[10] András Kappanyos, "Az egzakt irodalomtudomány kalandja" ["The Adventure of Quantitative Literary Studies"], *Alföld*, 63, 2012, 7, pp. 46-51.

phenomena, but also questions the legitimacy of this type of literary analysis in general. In other words, embodying the gap analysed above, he ignores the results of distant reading in the analysis of literary texts, and rejects quantitative methods in literary studies, whether based on manual or machine-based methods.

The beginnings and contemporary practice of computer-assisted distant reading have been strongly associated with the analysis of lyrical texts. Spanning nearly four decades, from the RPHA project initiated by Iván Horváth, to the ELTE Poetry Corpus[11], these efforts are less relevant to the present study's argument. One of the earliest attempts to undertake computational distant reading of novels, or at least to outline such a project, dates from the mid-2010's, and finds its theoretical foundation in Moretti's argument.

In his 2015 article, Gergely Labádi sketches a large-scale project aimed at creating a database of Hungarian novels encompassing not only the texts themselves but also their metadata and peritextual information. According to Labádi,

> Ideally, a database record should contain the precise text of the novel, along with structural and physical frameworks and functions related to the text (such as page numbers, chapters, mottos, etc.). Beyond the primary text, the record should also include peritextual elements (title page, dedication, engraving, advertisement, etc.). However, the database record does not merely transcribe these elements but organizes them into a schema adhering to the recommendations of TEI-XML, thus enriching them with metadata[12].

Although Labádi's ambitious plan to establish a comprehensive novel corpus was never realized due to his illness and untimely death, his conceptual framework anticipated several aspects of the research initiatives that have emerged in recent years concerning the computational distant reading of Hungarian novels. These aspects include the integrated handling of metadata and novel text on a single platform, which allows for both bibliographic and statistical linguistic analysis. Perhaps even more significantly, Labádi envisioned describing data in accordance with the guidelines of the Text Encoding Initiative (TEI), thereby providing a structured, metadata-enriched approach with a potential of international comparison.

Interestingly, the first Hungarian "novel database" was not established according to Labádi's proposal, nor was it initiated by Hungarian researchers. Instead, it emerged as part of an international project. Studies that analyse

[11] See https://github.com/ELTE-DH/poetry-corpus. Accessed March 12, 2025. See also the papers by Péter Horváth, e.g., Péter Horváth, "Két eljárás Magyar Nyelvű Versek metrumának gépi felismertetéséhez" ["Two Computational Methods for Detecting Meter in Hungarian Poetry"], *Digitális Bölcsészet*, 2021, 4, pp. 79-103.

[12] Gergely Labádi, "A magyar regény adatbázisa" ["The Database of the Hungarian Novels"], *Acta Historiae Litterarum Hungaricarum: Acta Universitatis Szegediensis,* 2016, 1, pp. 11-30. Unless otherwise stated, the quotations are translated into English by the author of this paper.

Hungarian literary history through computational linguistic analysis of large numbers of novels have relied on this corpus[13]. The project that fundamentally enabled these studies bears the highly evocative title *Distant Reading for European Literary History*.

Distant Reading for European Literary History (COST Action CA16204) is a project aiming to create a vibrant and diverse network of researchers jointly developing the resources and methods necessary to change the way European literary history is written. Grounded in the Distant Reading paradigm (i.e. using computational methods of analysis for large collections of literary texts), the Action will create a shared theoretical and practical framework to enable innovative, sophisticated, data-driven, computational methods of literary text analysis across at least 10 European languages. Fostering insight into cross-national, large-scale patterns and evolutions across European literary traditions, the Action will facilitate the creation of a broader, more inclusive and better-grounded account of European literary history and cultural identity[14].

The ambitious claim – reminiscent of Moretti's concept of "distant reading"[15] – that the methodology of literary historiography itself will change under the influence of CLS practice, remains difficult to verify, especially given the widening rather than narrowing gap between traditional and digital methodologies. However, it is undeniable that the multilingual novel corpus created within the framework of this project can reveal patterns accessible only through the perspective of distant reading, patterns that remain invisible to analogue methods and/or close reading.

The present study undertakes to present a previous scholarly experiment of reading novels "distantly", reproduce its results, extend its analysis, and evaluate the new findings thus obtained.

The Concept of the "Inward Turn"

The article titled "Towards a Computational History of Modernism in European Literary History: Mapping the Inner Lives of Characters in the European Novel, 1840–1920"[16] aims to investigate the inner worlds of characters through statistical linguistic tools, drawing upon Erich Kahler's concept of the "inward turn". The

[13] E.g. Botond Szemes, "A Sentence-Based Stylistic History of the Hungarian Novel", *Journal of Computational Literary Studies*, 2, 2023, 1, pp. 1-25.

[14] See the project website, https://www.distant-reading.net/ Accessed March 12, 2025.

[15] See Moretti, "Conjectures". While Moretti explicitly references Goethe's concept of world literature, he thereby seeks to elevate his own project – at least rhetorically – to Goethean heights.

[16] Tamara Radak, Lou Burnard, Pieter Francois et al., "Towards a Computational History of Modernism in European Literary History: Mapping the Inner Lives of Characters in the European Novel (1840–1920) [version 2; peer review: 2 approved, 1 approved with reservations], *Open Research Europe*, 2024, 3, 128, https://doi.org/10.12688/openreseurope.16290.2. Accessed March 12, 2025.

study attempts to derive insights regarding European literary history and the evolution of modernist prose from this quantitative approach. Although Kahler's influential essay from the late 1950s traces the "inward turn" only up to the end of the eighteenth century – precisely until the publication of *Tristram Shandy*[17], where the process reaches its peak and conclusion – he addresses the development of the European novel in the nineteenth and twentieth centuries in other essays. As R. Robertson summarizes it:

> The German émigré Erich Kahler composed a well-known study, *The Inward Turn of Narrative*, which traces the internalization of the novel from antiquity down to the eighteenth century. He pursued the topic in relation to modernism in a lecture given in 1958 and published as an appendix to his study, entitled "The Transformation of the Novel". Here Kahler pursues the inward turn through a number of stages. In the late nineteenth century, he argues, literature moved in one of two directions. One direction was collective: people were presented not primarily as isolated individuals but as members of a group, a crowd, a family, a social class, as in Zola's naturalist fiction or in Mann's Buddenbrooks where the individual is subordinate to the family. If collectivism looked outward toward society, the other direction moved inward, into the depths of the individual psyche. The visible surface of reality was decomposed to permit a minute analysis of sensations, emotions, psychological nuances[18].

According to Kahler's argument, the "inward turn" and the "outward turn" emerge as parallel processes at the beginning of the twentieth century, although the "outward turn" never became a commonplace concept in literary history, like its counter-notion. This tension is addressed by Melanie Conroy in her 2014 article[19], which effectively balances the fault line between quantitative and traditional literary analysis. Her study, which focuses on French prose history, serves as a significant precursor to both Radak et al. and the present study:

> One of the most common but seldom tested presuppositions about the alleged "inward turn" is that the linguistic innovations of the modernist period helped portray the mental states of characters in a more advanced manner. This article reviews the debates on the differences between realist and modernist forms of thought

[17] Erich Kahler, *The Inward Turn of Narrative* (1973). Transl. from the German by Richard and Clara Winston, Princeton, Princeton University Press, 2017, pp. 6-7: "For numerous reasons, personal as well as objective, our analysis stops at the end of the eighteenth century". Zoltán Abádi Nagy, one of the few Hungarian scholars referencing Kahler's literary-historical ideas, alludes precisely to this argument in connection with minimalist prose. See Zoltán Abádi Nagy, "Minimalizmus és narratív technika" ["Minimalism and Narrative Technique"], *Irodalomtörténet*, 1993, 1-2, p. 312. Kahler is not frequently cited in Hungarian literary history; his works are more often referenced in the context of national characterology.

[18] Ritchie Robertson, "Modernist Style and the 'Inward Turn' in German-Language Fiction", in Gregory Castle (ed.), *A History of the Modernist Novel*, New York and Cambridge, Cambridge University Press, 2015, pp. 293-310.

[19] Melanie Conroy, "Before the 'Inward Turn': Tracing Represented Thought in the French Novel (1800–1929)", *Poetics Today*, 35, 2014, 1-2, pp. 117-171.

> representation and uses quantitative techniques to determine whether, and to what extent, there was linguistic and stylistic innovation in how the French novel represented thought before and during the "inward turn" of the 1910's and 1920's.[20]

Conroy's conclusion is somewhat inconclusive; her quantitative examination of the linguistic representation of the "inward turn" does not reveal sharp trends or drastic shifts in the examined French novels. However, she identifies clear, albeit less radical, tendencies:

> The sharp increase in the frequency of mental verbs [...] provides strong evidence for the softer version of the modernist narrative advanced by critics like Lewis (2007), Matz (2006), and Schoenbach (2011). Accordingly, there is a gradual rise in interest in the mental states of characters through the nineteenth and early twentieth centuries[21].

When Radak et al. attempted similar investigations on a multilingual corpus in the above-mentioned study, they inevitably faced the fact that even within a single literary culture, there is no consensus on whether modernism (or its "opposite", realist prose) should be identified as a literary-historical period, an aesthetic category, or a narrative style (Conroy favours the former). Moreover, moving from the centre (e.g. English, German, French literature) to the periphery (such as Hungarian or Portuguese literature) – to invoke the contested centre-periphery model from contemporary discourse on modernism[22] – the temporal configurations of "literary movements" can take remarkably different forms. Indeed, it remains debatable whether such a comparison contributes productively to the literary-historical interpretation of a "national literature" at all[23].

Nevertheless, this does not diminish the significance of employing computational tools for linguistic statistical analysis on large, multilingual literary corpora. However, as Radak et al. themselves acknowledge, the question remains whether an analysis comparing the texts of only 100 novels per language can effectively reveal relevant correlations and patterns.

In light of the above, the following sections will briefly outline the methodology employed by Radak et al. in examining the novel corpus and then present the twofold extension of that research. This expansion involves the applying the methodology to a much larger Hungarian corpus, both in terms of time span and scope, compared to the original study conducted on 100 novels only.

[20] *Ibidem*, p. 117.

[21] *Ibidem*, p. 165.

[22] See Laura Winkiel, "The Modernist Novel in the World-System", in Gregory Castle (ed.), *A History*, pp. 408-428.

[23] Mihály Szegedy-Maszák, one of the few internationally known figures of Hungarian comparative literary studies, makes a sharp observation ("Comparative investigation often distorts national literature"), challenging the rhetoric of periodization in Hans Robert Jauss's *Studien zum Epochenwandel der ästhetischen Moderne*. See Mihály Szegedy-Maszák, "A kánonok szerepe az összehasonlító kutatásokban" ["The Comparative Analysis of Canons"], *Irodalomtörténet*, 76, 1995, 1, p. 7.

Examining Hungarian Novels: The Corpora

The investigations conducted by Radak et al. on Hungarian novels were based on 100 novels from the ELTeC corpus. The limited number of novels raises concerns about the generalizability of the findings to the broader historical processes of Hungarian literary history. Additionally, the logic underlying the construction of the corpus significantly influences the validity of the results.

The ELTeC corpus was assembled according to strict principles, and the composition of the Hungarian sub-corpus fully adhered to these guidelines. It is worth highlighting that these principles were not developed with the historical specificity of Hungarian literature (or other "peripheral" literature on the map of modernist fiction) in mind, nor were they based on literary-historical arguments. Rather, they were designed to ensure comparability between languages and the commensurability of sub-corpora. While it is impossible to quantify which sub-corpus or national literature is better or worse served by these compilation criteria, some aspects of representativeness can be addressed concerning the Hungarian sub-corpus. The goal of this study, however, is not to offer a critical analysis of previous quantitative research conducted on the ELTeC corpus, but to expand it. Radak et al. applied a similar method when they examined the relative frequency of certain verbs in an expanded corpus of French novels.

The compilation principles of the ELTeC corpus will be examined not in isolation, but in relation to the original Hungarian-language corpus and the expanded corpora utilized for this study. Accordingly, the remainder of this study analyses three distinct corpora in parallel.

The first corpus, ELTeC-hun, fully complies with the principles established by the *Distant Reading* project. It contains the texts and metadata of 100 novels, all of which are in the public domain and freely accessible via the project's GitHub repository. As required by the project's standards, this corpus includes only public domain texts.

The second corpus, ELTE Novel Corpus II, is being continuously developed by the Department of Digital Humanities at Eötvös Loránd University[24]. Like ELTeC-hun, it consists exclusively of public domain works, follows the same principles of markup and metadata encoding, and is freely accessible. This expanded corpus includes an additional 300 novels, incorporating all novels from ELTeC-hun I within its dataset. The 400 novels are sourced from the Hungarian Electronic Library (Magyar Elektronikus Könyvtár) in PDF, RTF, or HTML format, with two novels obtained from the Google Books service. The text layer of these works originates from the digital sources, comprising both corrected and uncorrected

[24] See ELTE Novel Corpus, https://github.com/ELTE-DH/regenykorpusz. Accessed March 12, 2025.

OCR results. During corpus construction, due to the lack of human resources, no efforts were made to correct OCR errors.

The third corpus, Novel Corpus III, diverges significantly from the previous two in that it also includes copyrighted texts, making it unavailable for public distribution. This corpus contains all novels from ELTE Novel Corpus II while also incorporating approximately 900 additional texts sourced from various repositories, including the Internet Archive, Google Books, the Digital Literary Academy (Digitális Irodalmi Akadémia), Project Gutenberg, and other Hungarian and international databases, as well as e-books.

The novels were selected based on their affiliation with the literary canon as suggested by high school textbooks, post-1989 literary history handbooks, and lists of literary prize winners. Due to the nature of text acquisition – employing web scraping technology where the base is an HTML code of a webpage rather than a document, or through the automated extraction of texts from formats such as EPUB using pre-existing tools – it is not possible to guarantee a constant high philological quality or completeness of the texts at the level of individual works[25]. The total size of the corpus is 1,297 works. The presumed date of first publication for each novel was assigned automatically through Python scripts using data from the MOKKA ODR service, the ISBN Database, and Wikidata. These dates were subsequently verified using ChatGPT (GPT-4o model) through a prompt including the name of the author and the title. In cases where publication dates from different sources showed discrepancies of more than five years, dates were manually corrected. When a novel was first published posthumously, the author's year of death was recorded instead of the publication date. During the semi-automated data correction process, numerous inaccuracies were also identified within the ELTE Novel Corpus II, which are currently being corrected. Since this corpus is versioned through a GitHub repository, all corrections will remain traceable.

The novels included in ELTeC-hun I and ELTE Novel Corpus II were annotated using the *emtsv pipeline* natural language processing (NLP) toolkit[26].

[25] See Fellegi, *A digitális filológia*, p. 132: "Moretti and the practice of distant reading in general have been frequently criticized for two main reasons. First, they do not strive for transparency in terms of their corpus and methodology, which raises concerns about reproducibility. Second, they often rely on texts of poor or questionable quality from a philological standpoint, which can also call their results into question". See also James Dobson, Scott Sanders, "Distant Approaches to the Printed Page", *Digital Studies/Le champ numérique*, 12, 2022, 1, pp. 1-28, and Katherine Bode, "The Equivalence of 'Close' and 'Distant' Reading; or, Toward a New Object for Data-Rich Literary History", *Modern Language Quarterly*, 2017, 1, pp. 77-106.

[26] See Balázs Indig, Bálint Sass, Eszter Simon, Iván Mittelholcz, Noémi Vadász, and Márton Makrai, "One Format to Rule Them All – The emtsv Pipeline for Hungarian", in Annemarie Friedrich, Deniz Zeyrek, and Jet Hoek (eds.), *Proceedings of the 13th Linguistic Annotation Workshop*, Florence, Association for Computational Linguistics, 2019, pp. 155-165, and Balázs Indig, Bálint Sass, and Iván Mittelholcz. "The xtsv Framework and the Twelve Virtues of Pipelines", in Nicoletta Calzolari et al. (eds.), *Proceedings of the Twelfth Language Resources and Evaluation Conference*, Marseille, European Language Resources Association, 2020, pp. 7044-7052.

The statistical linguistic findings presented in the study by Radak et al. are based on data tagged with part-of-speech (POS) annotations, represented in separate XML files following the "level2" annotation scheme. For the linguistic annotation of Novel Corpus III, we initially relied on the HuSpaCy library, utilizing its most accurate model built on the RoBERTa architecture[27]. However, the annotation quality – particularly in the case of punctuation and verbs – proved to be significantly weaker than that produced by the *emtsv* toolkit. Consequently, after applying whitespace normalization and deduplication across the entire corpus, we re-annotated the texts using the *emtsv pipeline* to ensure consistency and higher linguistic accuracy.

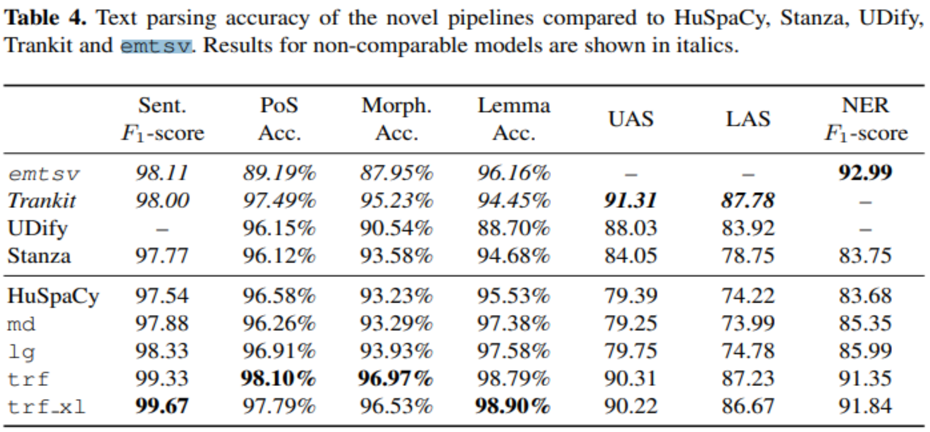


Table 4. Text parsing accuracy of the novel pipelines compared to HuSpaCy, Stanza, UDify, Trankit and emtsv. Results for non-comparable models are shown in italics.

	Sent.	PoS	Morph.	Lemma	UAS	LAS	NER
	F_1-score	Acc.	Acc.	Acc.			F_1-score
emtsv	*98.11*	*89.19%*	*87.95%*	*96.16%*	–	–	**92.99**
Trankit	*98.00*	*97.49%*	*95.23%*	*94.45%*	***91.31***	***87.78***	–
UDify	–	96.15%	90.54%	88.70%	88.03	83.92	–
Stanza	97.77	96.12%	93.58%	94.68%	84.05	78.75	83.75
HuSpaCy	97.54	96.58%	93.23%	95.53%	79.39	74.22	83.68
md	97.88	96.26%	93.29%	97.38%	79.25	73.99	85.35
lg	98.33	96.91%	93.93%	97.58%	79.75	74.78	85.99
trf	99.33	**98.10%**	**96.97%**	98.79%	90.31	87.23	91.35
trf_xl	**99.67**	97.79%	96.53%	**98.90%**	90.22	86.67	91.84

Divergence from ELTeC Criteria in the Corpora

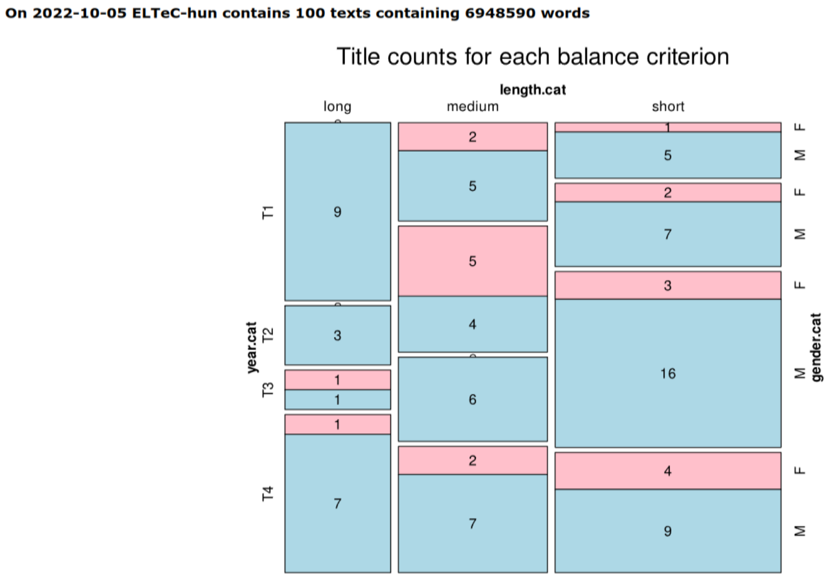
While we do not provide an exhaustive account of all ELTeC criteria here – these have been thoroughly documented in multiple studies and on the corpus's official GitHub repository[28] – we summarize the key principles, reflect on their potential implications, and highlight certain anomalies.

Since the principal motivation for creating the corpus was to uncover diachronic patterns and thereby contribute to a re-mapping of European literary history, the corpora were divided into four twenty-year periods (T1–T4). Each period was assigned a target number of works in order to ensure temporal balance.

[27] György Orosz et al., "Advancing Hungarian Text Processing with HuSpaCy: Efficient and Accurate NLP Pipelines", in Kamil Ekštein, František Pártl, Miloslav Konopík (eds.), *Text, Speech, and Dialogue. TSD 2023*, Cham, Springer, 2023, pp. 58-69.

[28] E.g. Lou Burnard, Christof Schöch and Carolin Odebrecht, "In Search of Comity: TEI for Distant Reading", *Journal of the Text Encoding Initiative*, April 2021–March 2023, 14, https://journals.openedition.org/jtei/3500. Accessed March 12, 2025.

In the case of each national ELTeC subcorpus, the aim was to achieve balance across several parameters within each time segment: text length (long, medium, and short works), author gender, and canonical status (as indicated by the frequency of reprints). The following figure visualizes the application of these criteria within the Hungarian sub-corpus of 100 novels.



On 2022-10-05 ELTeC-hun contains 100 texts containing 6948590 words

Title counts for each balance criterion

Source: GitHub ELTeC repository (https://distantreading.github.io/ELTeC/hun/index.html)

The ELTE Novel Corpus II adheres to the ELTeC corpus in terms of markup and metadata structure but diverges from it in terms of corpus composition. Its purpose is not comparability with other corpora, but rather to support a broad range of statistical linguistic analyses and to expand the annotated text base underlying a search interface developed for the corpus[29].

Compared to ELTeC, the proportion of male authors and canonized works (as indicated by frequent reprints) is significantly higher. This skew reflects the canon-centric tendencies of digitization processes. The majority of texts in the corpus originate from the Hungarian Electronic Library (MEK), a service maintained by the National Széchényi Library[30]. MEK does not specify formal criteria for inclusion, nor does it prioritize first editions – likely for reasons related to conservation: protecting vulnerable old editions. Moreover, as an electronic library, MEK is oriented toward general readership needs, all of which contribute to the over-representation of canonized works in the corpus.

[29] See ELTE Novel Corpus search interface, https://regenykorpusz.elte-dh.hu/. Accessed March 12, 2025.

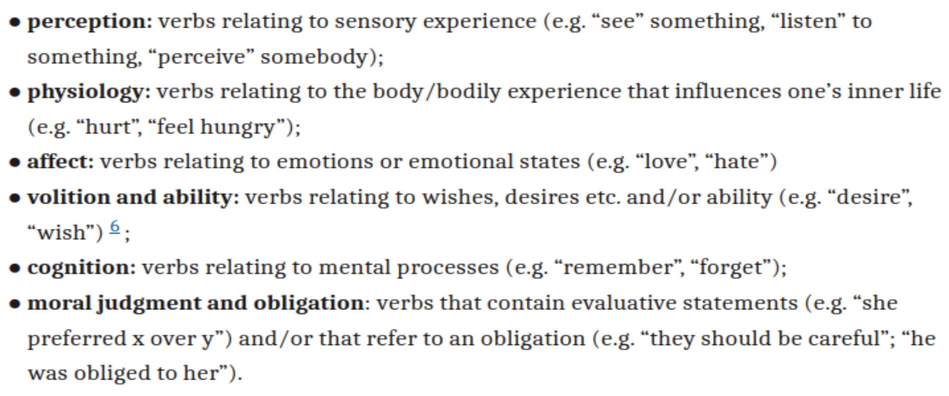
[30] See Magyar Elektronikus Könyvtár, https://www.mek.oszk.hu/hu/aboutus/#introduction. Accessed March 12, 2025.

In the case of Novel Corpus III, the tendencies described above are even more pronounced: the patterns of the Hungarian literary canon – most notably the disproportionate presence of male authors – are even more clearly reflected across all time periods. While the primary focus during the corpus's construction was to determine the first publication date of each novel, the tracking of subsequent reprints was not part of the process. This omission was intentional, as it would have significantly slowed corpus compilation, and our goal was to assemble a corpus of Hungarian novels comparable in size to the French corpus analysed by Radak et al.

The ELTeC corpora imposed strict limits on the number of works allowed from any single author – again, in the interest of preserving corpus diversity. However, this restriction was not applied in the construction of ELTE Novel Corpus II or III.

Relative Frequency of "Inner Life Verbs" as a Metric

Radak et al. proceed from the hypothesis that the "inward turn" – that is, the modernist novel's increasing focus on consciousness and the inner lives of characters – can be detected through changes in the frequency of certain verbs. Drawing on Bretherton and Beeghly's grouping of utterances relating to mental states[31], they derived six relevant categories that were used when compiling the list of verbs to be mapped:



- **perception:** verbs relating to sensory experience (e.g. "see" something, "listen" to something, "perceive" somebody);
- **physiology:** verbs relating to the body/bodily experience that influences one's inner life (e.g. "hurt", "feel hungry");
- **affect:** verbs relating to emotions or emotional states (e.g. "love", "hate")
- **volition and ability:** verbs relating to wishes, desires etc. and/or ability (e.g. "desire", "wish") [6];
- **cognition:** verbs relating to mental processes (e.g. "remember", "forget");
- **moral judgment and obligation**: verbs that contain evaluative statements (e.g. "she preferred x over y") and/or that refer to an obligation (e.g. "they should be careful"; "he was obliged to her").

Source: Radak et al., p. 5.

The verbs used to measure the frequency of *inner-life verbs* in the Hungarian corpus were categorized according to semantic domains, following the methodology of Radak et al. The classification is as follows:

[31] Inge Bretherton and Marjorie Beeghly, "Talking about Internal States: The Acquisition of an Explicit Theory of Mind", *Developmental Psychology*, 18, 1982, 6, pp. 906-921.

- **Perception**
  - *lát* (“sees”)
  - *néz* (“looks”)
  - *hall* (“hears”)
- **Affect**
  - *szeret* (“loves”)
  - *érez* (“feels”)
  - *tetszik* (“likes” / “pleases”)
- **Volition**
  - *akar* (“wants”)
  - *kíván* (“desires”)
  - *remél* (“hopes”)
- **Cognition**
  - *tud* (“knows” / “can”)
  - *gondol* (“thinks”)
  - *ért* (“understands”)
- **Moral Judgement / Obligation**
  - *kell* (“must”)
  - *megenged* (“permits”)
  - *engedelmeskedik* (“obeys”)

Based on their findings, Radak et al. concluded that a slight upward trend in the frequency of *inner-life verbs* could be observed in both the Hungarian and English corpora. This observation holds when the trend calculation is based on the relative frequencies of these verbs within individual novels, treating each text as a separate data point. However, because the frequencies vary considerably even within a single decade, it is methodologically more sound – especially in light of the corpus’s temporal segmentation into four periods (T1–T4) – to assess the change in aggregated relative frequencies across broader time intervals.

When viewed in this way, the increase in the frequency of *inner-life verbs* becomes much more apparent. In the case of the 100 novels in ELTeC-hun, the ratio of *inner-life verbs* to all verbs rises from 9.17% in 1840 to 11.5% in 1910. A similar trend can be observed in the ELTE Novel Corpus II, where the measured increase over a slightly longer period (1840–1930) is even more pronounced, with frequencies rising from 8.98% to 11.5% across 400 novels.

A parallel trend is observable when examining the ratio of verbs to all words in both corpora – the 100-novel ELTeC-hun and the 400-novel ELTE Novel Corpus. This raises the question of whether the observed increase in verb frequency might be an artifact of the methodology or a limitation of the natural language processing (NLP) tools used in the analysis.

The methodology adopted by Radak et al. closely follows the specifications developed within the *Distant Reading* project. During the creation of the so-called “level2” markup, NLP tools were employed to automatically segment the novel texts into sentences, then into words and punctuation marks. Words were assigned part-of-speech (POS) tags and lemmatized. Verbs not recognized as such by the POS tagger were, by default, excluded from both the Radak et al. analysis and the frequency calculations presented above.

However, the NLP tools available for Hungarian – such as the *emtsv pipeline* used in all three phases of this research – are optimized for processing contemporary Hungarian texts. As a result, their performance is less reliable when applied to 19th-century novels, particularly in relation to morphological phenomena that have undergone significant change over time. The observed increase in verb frequency during the 19th century thus raises the suspicion that some of the trend may be attributable to tagging errors, especially within the POS tagging process.

In the history of Hungarian verb morphology, perhaps the most significant change concerns the transformation of the verbal tense system:

> The writers and poets of the Enlightenment and Reform Era undertook the task of language renewal with unprecedented determination and effectiveness. This activity left its most visible imprint […] on the system of verb suffixes. From an aesthetic perspective, driven by a desire for variety, authors reintroduced into literary usage the undeniably rich system of verbal tenses, which, in spoken language, had been gradually simplifying since the sixteenth century. Both the narrative past [*elbeszélő múlt]* and the simple past tense formed with *-t* were used, even though there had long ceased to be a functional distinction between them. At the same time, because the narrative past had almost entirely disappeared from colloquial usage, it gradually acquired a more elevated, literary stylistic value compared to the other past tense[32].

Within the discourse of Hungarian historical linguistics and literary stylistics, there is broad agreement that the literary tradition of the 19th century made extensive use of past tense verb forms that had already fallen out of use in spoken language. However, scholars differ considerably in their interpretation of the motivations behind this practice. Some argue that it reflects an aesthetic striving for variation, solemnity, or an elevated literary tone – implying, as Zsófia Sárosi suggests in the above quotation, that there was no functional difference between the various past tenses in prose language.

By contrast, Gábor Tolcsvai Nagy, in his examination of 19th-century literary language, convincingly demonstrates that the *elbeszélő múlt* (*narrative past*) fulfilled a distinct function within the literature of the period, contributing to the layered representation of the past in fictional narrative. Through the analysis of 19th-century diary texts, he quantitatively substantiates the importance of the *narrative past tense* in literary narrative[33].

[32] Zsófia Sárosi, "Morfématörténet" ["The Historical Development of Morphemes"], in Jenő Kiss, Ferenc Pusztai (eds)., *Magyar nyelvtörténet. A magyar nyelv története / Az újmagyar kor* [*History of the Hungarian Language. The Early Modern Period*], Budapest, Osiris, 2003, p. 131.

[33] Gábor Tolcsvai Nagy, "Mondá – mondta. A folyamatos és az egyszerű múlt idő konstruálási mintázatai 19. századi naplókban" ["Mondá – mondta. Patterns in the Use of Narrative and Simple Past Tenses in the 19th-Century Diaries"], *Magyar Nyelvőr*, 2021, 4, pp. 432-447.

For the purposes of the present study, however, it is not essential to determine the aesthetic or stylistic function of the narrative past in the literature of the period under investigation. What is far more relevant is the fact that the NLP tools used to annotate the corpus exhibit a remarkably high error rate in tagging these verb forms correctly. To demonstrate this, I deviated from the workflow employed by the ELTeC project and Radak et al., identifying the narrative past tense verb forms of *inner-life verbs* at the token level rather than the lemma level. This approach allowed me to bypass the inaccuracies introduced by the POS tagger. However, it also necessitated the exclusion of homonymous forms from the analysis. This affected three word forms:

> *gondol (verb) - gondola* (past tense verb or noun meaning Venetian boat)
> *ért (verb) - érte* (past tense verb or pronoun meaning "for her/him")
> *érez (verb) - érzék* (past tense verb or noun meaning "sense").

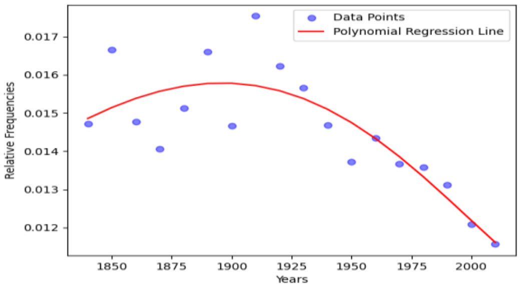
The key question, however, is how this extended verb identification affects the previously observed upward trend in the 100- and 400-novel corpora. The results are unambiguous: the previously identified increase of 2–2.5% in *inner-life verb* frequency disappears. For the 100-novel corpus, the increase falls below 1%; for the 400-novel corpus, it drops below 1.2%. Surprisingly, in the 1,200-novel corpus, the standard deviation becomes even smaller.

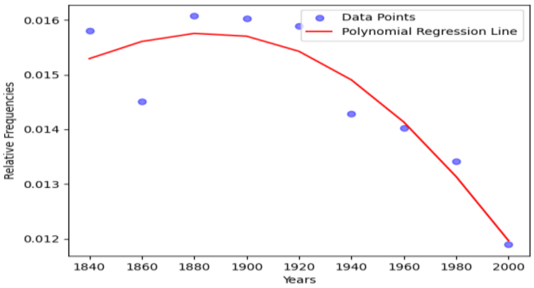
Despite the significant improvements in both methodology and corpus scale, these findings are, in essence, consistent with the conclusions drawn by Radak et al. – namely, that the phenomenon of the "inward turn" cannot be reliably identified through the statistical analysis of the relative frequency of *inner-life verbs*, as their proportion, at least for Hungarian, remains remarkably stable over the time-span examined.

Nevertheless, the study – while not elaborating on it in detail – does point toward a possible direction for further inquiry: a relative frequency of the *affect* category of *inner-life verbs* in the extended French data (1750–2000) shows a visibly decreasing trend. Building on this observation, the present study now turns to an analysis of potential patterns in the frequency changes of specific verb subgroups, categorized according to the typology proposed by Bretherton and Beeghly.

Trends in Subcategories of Inner-Life Verbs in the Novel Corpus III

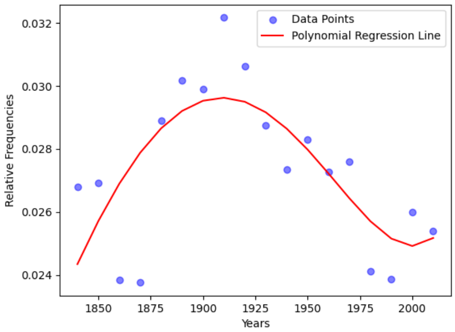
As previously noted, the overall relative frequency of *inner-life verbs* across twenty-year periods shows only minor fluctuations – variations amounting to mere fractions of a percent. In contrast, verbs associated with *affective states* exhibit a much more distinct and pronounced pattern.





In both approaches – whether examining relative frequencies across single decades or twenty-year intervals – the patterns can be considered statistically significant. This marks the first pattern in the data that may plausibly be linked to the hypothesized shift associated with turn-of-the-century modernity: the relative frequency of the three most common affective verbs peaks around 1900. Even more striking is the observation that the relative frequency of these verbs subsequently declines in a consistent downward trend, a trend that requires further interpretation.

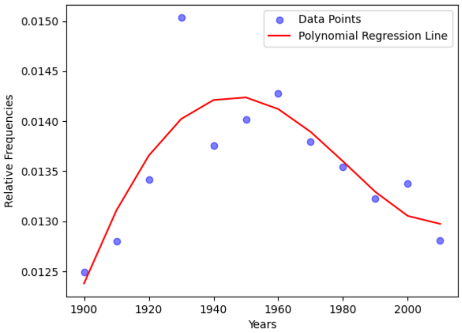
The next verb group to exhibit a statistically significant pattern consists of the most frequent verbs associated with *perception*.



The frequency of verbs denoting *sensory perception* peaks significantly at the turn of the 20th century, surpassing the values observed in both earlier 19th-century literature and in literary language from the 20th century to the present. The trend in this verb group parallels that of the *affect verbs* in two important respects: not only do both indicate a statistically traceable shift around 1900, but they also share the characteristic that no other comparable "period shifts" – in the sense of a distinct reversal or transformation – can be identified elsewhere in the timeline. While the sensory verbs seem to show a turning point at the turn of the millennium, the variation within individual decades is too great to draw firm conclusions.

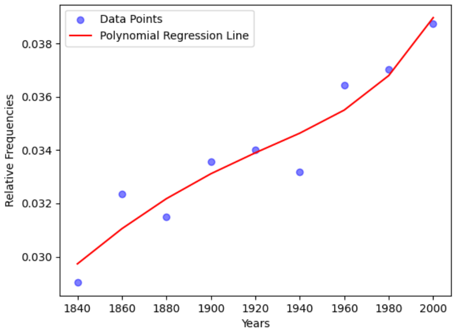
The third verb group to display a noteworthy pattern is that associated with *volition*. Although the frequency change in *volition-related verbs* is not statistically significant when considering the full corpus, a compelling pattern emerges when the data is restricted to the period between 1890 and 2010. Within this time frame, not only does the trend reach statistical significance, but it also aligns intriguingly with broader historical and cultural shifts.

The relative frequency of *volition verbs* in the corpus increases sharply within a short time span, with the upward trend appearing to reverse sometime after the 1960s. Particularly striking is the fact that the lowest frequency is measured at the turn of the century, while novels from the 1920s and 1930s show markedly elevated frequencies for this verb group.



The final verb group displaying a potentially meaningful pattern in frequency change is associated with *cognition*. Somewhat unexpectedly, this group exhibits a consistently increasing trend across the entire corpus. In this regard, it appears least relevant to the broader aims of the study, as no clear inflection points or historical shifts can be identified.

Nevertheless, a nearly 10% difference in relative frequency is observable between the earliest and the most recent 20 years periods in the corpus. While this finding does not support hypotheses regarding specific temporal breaks or turning points, it suggests an underlying long-term trend that warrants further investigation.



Conclusion: The Frequency of Inner-Life Verbs and Hungarian Literary Modernism

This study set out to examine whether a concrete practice of computational reading – developed within the framework of the *Distant Reading for European Literary History* project and centred on a comparative literary-historical analysis of the multilingual ELTeC novel corpus – might yield more meaningful results when applied to a significantly larger corpus of novels. The original application of this methodology produced only limited findings of literary-historical relevance. From this perspective, the outcome of our investigation is unambiguously negative: applying the methodology of Radak et al. to a POS-tagged corpus of 400 or even 1,200 Hungarian novels reveals no patterns that could be interpreted as meaningful from the standpoint of literary history.

At the same time, the analysis was able to demonstrate that in the statistical linguistic analysis of 19th-century Hungarian literature, the high error rate of POS tagging for *narrative past tense* verb forms introduces a measurable distortion. It can also be stated with confidence that once these tagging errors are corrected using our proposed method, the proportion of inner-life verbs – as defined in Radak et al. – remains surprisingly consistent from the early 1800s to contemporary fiction.

Following this, our analysis focused on the frequency shifts within smaller subgroups of the inner-life verb set. Two of these subgroups displayed statistically significant patterns: both the *perception* and *affect* verb groups peaked around 1900, a finding that aligns closely with the most widely accepted literary-historical interpretations of Hungarian modernism[34].

Moreover, the linguistic data also reflects two further literary paradigm shifts widely recognized within Hungarian literary history. The *volition* verb group shows a marked increase, reaching a local maximum in novels from the 1920s and

[34] See Tibor Gintli (ed.), *Magyar irodalom* [*Hungarian Literature*], Budapest, Akadémiai Kiadó, 2010.

1930s. This change corresponds to what is often described as the period of second modernism – a post-avant-garde phase of literary renewal. This group then begins a continuous decline in the 1960s and 1970s, roughly preceding what is known as the prose turn: a linguistic and conceptual transformation in Hungarian prose fiction from the 1970s onward. One of the landmark achievements of this period is the appearance of Péter Esterházy's *Termelési-regény* [*A Production Novel*] in the late 1970's, considered symbolically significant alongside the works of the so-called "generation of the Péters"[35].

Can this be taken to mean that the methodology of CLS – promising, in the vision of Moretti and later of the *Distant Reading* project, to uncover new patterns in literary history – indeed fulfils this function? Hardly. While we have demonstrated that shifts in the relative frequency of just a handful of common verbs align with familiar literary-historical periodizations, the method is far from sophisticated enough to provide substantial arguments either in support of or against such periodizations, let alone to revise widely accepted historical boundaries.

However, the verb frequency data does reveal other types of patterns. When we examine the minimum and maximum values of verb frequency at the level of individual texts, we encounter unexpected constellations that demand explanation. Some cases are relatively straightforward. For example, it is no surprise that Péter Esterházy's *A Woman* ranks among the "winners" in the frequency of *affect verbs*. The short volume, composed of brief prose sequences, builds its poetic rhythm on the repetition of minimalist, fragmentary sentences such as "There is a woman", "She loves", and "She hates". It is much more difficult to explain why several of Iván Mándy's works appear at the bottom of the frequency list for affect verbs – where these verbs make up only 6% of the total verbs count – while many of Lajos Kassák's works exceed the 20% mark, which is a threefold difference and puts him at the top of the same list.

Intriguing contrasts also emerge among the most influential works of the so-called "generation of the Péters". *A Production Novel*, for instance, contains an exceptionally low proportion of *volition verbs,* the lowest in Esterházy's oeuvre, and similarly low values are seen in Péter Lengyel's iconic *Macskakő* [*Cobblestone*]. This stands in stark contrast to Péter Nádas's *Emlékiratok könyve* [*A book of memories*], which registers 3 times more *volition verbs*. Such findings further weaken the already tenuous claim that post-1970s changes in *volition verb* frequency correspond to the renewal of Hungarian prose.

The closer analysis of these and similar questions lies beyond the scope of the present study. My aim here has merely been to point out that even this relatively simple application of CLS technologies – analyzing verb frequency through a

[35] See Ágnes Balajthy, Katalin Bódi, Péter Szirák, *A kortárs magyar irodalom* [*Contemporary Hungarian Literature*], Debrecen, Debrecen University Press, 2021.

literary-historical lens – may still yield unexpected insights and reveal patterns that challenge and productively provoke the conventions of traditional literary-historical discourse. But in order for CLS to propose new temporal patterns in place of, or alongside, the existing literary historical periodization, and to convince sceptical traditional literary scholars, a more complex methodology than the above is required. Vector space semantics[36] offers itself as such a methodology, but that is a topic for another paper.

BIBLIOGRAPHY

ABÁDI NAGY, Zoltán, "Minimalizmus és narratív technika" ["Minimalism and Narrative Technique"], *Irodalomtörténet*, 1993, 1-2, pp. 311-323.

BALAJTHY, Ágnes, BÓDI, Katalin, SZIRÁK, Péter, *A kortárs magyar irodalom* [*Contemporary Hungarian Literature*], Debrecen, Debrecen University Press, 2021.

BRETHERTON, Inge, BEEGHLY, Marjorie, "Talking about Internal States: The Acquisition of an Explicit Theory of Mind", *Developmental Psychology*, 18, 1982, 6, pp. 906-921.

BODE, Katherine, "The Equivalence of 'Close' and 'Distant' Reading; or, Toward a New Object for Data-Rich Literary History", *Modern Language Quarterly*, 2017, 1, pp. 77-106.

BURNARD Lou, SCHÖCH, Christof, ODEBRECHT, Carolin, "In Search of Comity: TEI for Distant Reading", *Journal of the Text Encoding Initiative*, April 2021–March 2023, 14, https://journals.openedition.org/jtei/3500. Accessed March 12, 2025.

CIOTTI, Fabio, "Distant Reading in Literary Studies: A Methodology in Quest of Theory", *Testo e Senso*, 2021, 23, pp. 195-213.

CONROY, Melanie, "Before the 'Inward Turn': Tracing Represented Thought in the French Novel (1800–1929)", *Poetics Today*, 35, 2014, 1-2, pp. 117-171.

DA, Nan Z., "The Computational Case against Computational Literary Studies", *Critical Inquiry*, 45, 2019, 3, pp. 601-639.

DOBSON, James E., "Vector Hermeneutics: On the Interpretation of Vector Space Models of Text", *Digital Scholarship in the Humanities*, 2022, 1, pp. 81-93.

DOBSON, James, SANDERS, Scott, "Distant Approaches to the Printed Page", *Digital Studies/Le champ numérique*, 12, 2022, 1, pp. 1-28.

FELLEGI, Zsófia, *A digitális filológia Magyarországon: elvek és gyakorlatok* [*Digital Philology in Hungary: Principles and Practices*], Doctoral Thesis, Pázmány Péter Catholic University, Budapest, 2024, p. 17, https://disszertacio.ppke.hu/id/eprint/583/. Accessed March 12, 2025.

GINTLI, Tibor (ed.), *Magyar irodalom* [*Hungarian Literature*], Budapest, Akadémiai Kiadó, 2010.

HORVÁTH Péter et al., "ELTE Verskorpusz: a magyar kanonikus költészet gépileg annotált adatbázisa" ["ELTE Poetry Corpus: A Machine Annotated Database of Canonical Hungarian Poetry"], in Gábor Berend et al. (eds.), *XVIII. Magyar Számítógépes Nyelvészeti Konferencia*

[36] See Andrey Kutuzov, Lilja Øvrelid, Terrence Szymanski, Erik Velldal, "Diachronic Word Embeddings and Semantic Shifts: A Survey", in Emily M. Bender, Leon Derczynski, and Pierre Isabelle (eds.), *Proceedings of the 27th International Conference on Computational Linguistics (COLING 2018)*, Santa Fe, Association for Computational Linguistics, 2018, pp. 1384-1397; James E. Dobson, "Vector Hermeneutics: On the Interpretation of Vector Space Models of Text", *Digital Scholarship in the Humanities*, 2022, 1, pp. 81-93.

(MSZNY 2022) [*18th Hungarian Conference on Computational Linguistics*], Szeged, JATEPress, 2022, pp. 375-388.

HORVÁTH, Péter, "Két eljárás Magyar Nyelvű Versek metrumának gépi felismertetéséhez" ["Two Computational Methods for Detecting Meter in Hungarian Poetry"], *Digitális Bölcsészet*, 2021, 4, pp. 79-103.

INDIG, Balázs, SASS, Bálint, MITTELHOLCZ, Iván, "The xtsv Framework and the Twelve Virtues of Pipelines", in Nicoletta Calzolari et al. (eds.), *Proceedings of the Twelfth Language Resources and Evaluation Conference*, Marseille, European Language Resources Association, 2020, pp. 7044-7052.

INDIG, Balázs, SASS, Bálint, SIMON, Eszter, MITTELHOLCZ, Iván, VADÁSZ, Noémi, MAKRAI, Márton, "One Format to Rule Them All – The emtsv Pipeline for Hungarian", in Annemarie Friedrich, Deniz Zeyrek, and Jet Hoek (eds.), *Proceedings of the 13th Linguistic Annotation Workshop*, Florence, Association for Computational Linguistics, 2019, pp. 155-165.

KAHLER, Erich, *The Inward Turn of Narrative* (1973). Transl. from the German by Richard and Clara Winston, Princeton, Princeton University Press, 2017.

KAPPANYOS András, "Az egzakt irodalomtudomány kalandja" ["The Adventure of Quantitative Literary Studies"], *Alföld*, 63, 2012, 7, pp. 46-51.

KUTUZOV, Andrey, ØVRELID, Lilja, SZYMANSKI, Terrence, VELLDAL, Erik, ""Diachronic Word Embeddings and Semantic Shifts: A Survey", in Emily M. Bender, Leon Derczynski, and Pierre Isabelle (eds.), *Proceedings of the 27th International Conference on Computational Linguistics (COLING 2018)*, Santa Fe, Association for Computational Linguistics, 2018, pp. 1384-1397.

LABÁDI, Gergely, "A magyar regény adatbázisa" ["The Database of the Hungarian Novels"], *Acta Historiae Litterarum Hungaricarum: Acta Universitatis Szegediensis,* 2016, 1, pp. 11-30.

LUHMANN, Niklas, *Art as a Social System*. Transl. by Eva M. Knodt, Stanford, Stanford University Press, 2000.

MORETTI, Franco, "Conjectures on World Literature", *New Left Review*, 2000, 1, pp. 54-68.

OROSZ, György et al., "Advancing Hungarian Text Processing with HuSpaCy: Efficient and Accurate NLP Pipelines", in Kamil Ekštein, František Pártl, Miloslav Konopík (eds.), *Text, Speech, and Dialogue. TSD 2023*, Cham, Springer, 2023, pp. 58-69.

RADAK, Tamara, BURNARD, Lou, FRANCOIS, Pieter et al., "Towards a Computational History of Modernism in European Literary History: Mapping the Inner Lives of Characters in the European Novel (1840–1920) [version 2; peer review: 2 approved, 1 approved with reservations], *Open Research Europe*, 2024, 3, 128, https://doi.org/10.12688/openreseurope.-16290.2. Accessed March 12, 2025.

RIES, Thorsten, van DALEN-OSKAM, Karina, OFFERT, Fabian, "Reproducibility and Explainability in Digital Humanities", *International Journal of Digital Humanities*, 2024, 6, pp. 1-7.

ROBERTSON, Ritchie, "Modernist Style and the 'Inward Turn' in German-Language Fiction", in Gregory Castle (ed.), *A History of the Modernist Novel*, New York and Cambridge, Cambridge University Press, 2015, pp. 293-310.

SÁROSI Zsófia, "Morfématörténet" ["The Historical Development of Morphemes"], in Jenő Kiss, Ferenc Pusztai (eds)., *Magyar nyelvtörténet. A magyar nyelv története / Az újmagyar kor* [*History of the Hungarian Language. The Early Modern Period*], Budapest, Osiris, 2003, pp. 129-153.

SCHÖCH, Christof, DUDAR, Julia, FILEVA, Evgeniia, ŠEĻA, Artjoms, "Multilingual Stylometry: The Influence of Language on the Performance of Authorship Attribution using Corpora from the European Literary Text Collection (ELTeC)", in Wouter Haverals, Marijn Koolen, Laure Thompson (eds.), *Proceedings of the Computational Humanities Research Conference 2024*, Aachen, PublisherCEUR-WS.org, 2024, pp. 386-408.

SZEGEDY-MASZÁK, Mihály, "A kánonok szerepe az összehasonlító kutatásokban" ["The Comparative Analysis of Canons"], *Irodalomtörténet*, 76, 1995, 1, pp. 5-36.

SZEMES, Botond et al., "Az ELTE Drámakorpuszának létrehozása és lehetőségei" ["The Creation and Potential Applications of ELTE Drama Corpus"], in József Tick, Károly Kokas, András Holl (eds.), *Valós térben, Az online térért: Networkshop 31: országos konferencia* [*In Real Space, for Online Space: Networkshop 31 – National Conference*], Budapest, HUNGARNET Egyesület, pp. 170-178.

SZEMES Botond, "A Sentence-Based Stylistic History of the Hungarian Novel", *Journal of Computational Literary Studies*, 2, 2023, 1, pp. 1-25.

TOLCSVAI NAGY Gábor, "Mondá – mondta. A folyamatos és az egyszerű múlt idő konstruálási mintázatai 19. századi naplókban" ["Mondá – mondta. Patterns in the Use of Narrative and Simple Past Tenses in the 19th-Century Diaries"], *Magyar Nyelvőr*, 2021, 4, pp. 432-447.

WINKIEL, Laura, "The Modernist Novel in the World-System", in Gregory Castle (ed.) *A History of the Modernist Novel*, New York and Cambridge, Cambridge University Press, 2015, pp. 408-428.

THE RELATIVE FREQUENCY OF INNER-LIFE VERBS AS SIGNIFIER OF CHANGE IN 19TH AND 20TH CENTURY FICTION? DISTANT READING OF A CORPUS OF HUNGARIAN NOVELS
(*Abstract*)

This study reexamines a key hypothesis of computational literary studies (CLS): that long-term historical shifts in narrative style – particularly the so-called "inward turn" associated with modernist fiction – can be detected through the relative frequency of verbs related to mental and emotional states ("inner-life verbs"). Building on the methodology proposed by Radak et al. within the *Distant Reading for European Literary History* project, this research significantly expands both the corpus and the linguistic precision of the analysis. While the original study was based on 100 Hungarian novels from the ELTeC corpus, we extend the analysis to over 1,200 Hungarian novels spanning the 19th to the 21st century, annotated with enhanced natural language processing tools and corrected for systematic tagging errors – particularly those affecting past tense verb forms in historical prose. Our findings suggest that the overall frequency of inner-life verbs remains strikingly stable over the two centuries studied, undermining claims that this metric can effectively signal major literary-historical transitions. However, by examining semantic subgroups of inner-life verbs, we identify statistically significant trends that align with established periodizations in Hungarian literary history. Specifically, verbs of perception and affect peak around 1900, while volition-related verbs rise sharply between the 1920s and 1930s before declining from the 1960s onward – mirroring the "second modernism" and the later prose turn of the 1970s. At the same time, the growing prevalence of cognition-related verbs suggests a long-term, less period-specific trend. Although our findings confirm that inner-life verb frequency alone is insufficient to redefine literary periodization, they do reveal intriguing patterns at the level of individual works and authors, calling for further investigation. The study thus demonstrates both the limitations and the untapped potential of CLS methodologies. While statistical verb analysis may not yet replace traditional literary-historical reasoning, it can productively complement and challenge it. For CLS to fulfil its promise of offering new historical insights, more sophisticated semantic models – such as diachronic vector space embeddings – must be integrated into future research.

Keywords: computational literary studies, distant reading, Hungarian novel corpus, inward-turn, periodization.

FRECVENȚA RELATIVĂ A VERBELOR CARE EXPRIMĂ INTERIORITATEA CA INDICATOR AL TRANSFORMĂRILOR ÎN PROZA SECOLELOR AL XIX-LEA ȘI AL XX-LEA? UN DEMERS DE DISTANT READING AL UNUI CORPUS DE ROMANE MAGHIARE (*Rezumat*)

Acest studiu revizitează o ipoteză centrală a studiilor literare computaționale (SLC): faptul că transformările stilului narativ de-a lungul unor etape istorice ample – în special așa-numita "întoarcere spre interior" asociată cu ficțiunea modernistă – pot fi detectate prin frecvența relativă a verbelor care exprimă stări mentale și emoționale ("verbe ale interiorității"). Bazându-se pe metodologia propusă de Radak et al. în cadrul proiectului *Distant Reading for European Literary History*, această cercetare extinde semnificativ atât corpusul, cât și precizia lingvistică a analizei. În timp ce studiul amintit se baza pe 100 de romane maghiare cuprinse în corpusul ELTeC, noi extindem analiza la peste 1.200 de romane maghiare din secolele al XIX-lea și al XXI-lea, adnotate cu instrumente avansate de procesare a limbajului natural și corectate în ceea ce privește erorile sistematice de etichetare – în special cele care vizează formele de trecut ale verbelor utilizate în proza istorică. Rezultatele noastre sugerează că frecvența globală a verbelor ce exprimă interioritatea rămâne remarcabil de stabilă de-a lungul celor două secole analizate, punând sub semnul întrebării afirmațiile conform cărora acest indicator ar putea semnala eficient marile metamorfoze istorico-literare. Totuși, prin examinarea subgrupurilor semantice ale acestor verbe, pot fi delimitate tendințe semnificative statistic, care se corelează cu periodizările consacrate ale istoriei literare maghiare. Mai precis, verbele care exprimă percepția și afectivitatea sunt utilizate mai ales în jurul anului 1900, în timp ce verbele care exprimă voința cunoscu o frecvență mai însemnată între anii 1920 și 1930, pentru ca apoi să scadă după anii 1960 – fenomen ce reflectă așa-numitul "al doilea modernism" al literaturii din Ungaria și transformarea totală a prozei din anii 1970. În același timp, creșterea constantă a frecvenței verbelor care exprimă procese cognitive indică o tendință pe termen lung, mai puțin dependentă de perioade specifice. Deși rezultatele noastre confirmă faptul că simpla constatare a frecvenței verbelor interiorității nu este suficientă pentru a redefini periodizarea literară, ele relevă totuși o serie de tipare inedite care caracterizează opere și autori particulari, legitimând astfel utilitatea unor investigații suplimentare. Studiul demonstrează astfel atât limitele, cât și potențialul neexplorat al metodologiilor SLC. Chiar dacă analiza statistică a formelor verbale nu poate încă înlocui abordările istorico-literare tradiționale, ea le poate completa și revizui. Pentru ca SLC să-și împlinească promisiunea de a oferi perspective istorico-literare noi, viitoarele cercetări trebuie să integreze modele interpretative mai sofisticate – cum ar fi reprezentările vectoriale diacronice ale construcției semnificațiilor literare.

Cuvinte-cheie: studii literare computaționale, distant reading, corpus de romane maghiare, întoarcerea spre interioritate, periodizare literară.