


A Sentence-Based Stylistic History of the Hungarian Novel

Botond Szemes¹ 

1. Institute for Literary Studies, Research Centre for the Humanities, Budapest, Hungary.

Citation

Botond Szemes (2023). "A Sentence-Based Stylistic History of the Hungarian Novel". In: *Journal of Computational Literary Studies* 2 (1). 10.48694/jcls.3582

Date published 2024-02-10

Date accepted 2023-01-24

Date received 2023-02-17

Keywords

stylemetry, sentence structure, clause linkage, literary history, classification, epistemology

License

CC BY 4.0 

Reviewers

Anouk Lang, Robert Hesselbach

Note

This paper has passed through the conference track of JCLS. In addition to being peer reviewed, it was presented and discussed at the 2nd Annual Conference of Computational Literary Studies 2023 at the University of Würzburg.

Abstract. The paper presents a method for the automatic identification of different types of compound and complex sentences in Hungarian through the analysis of conjunctions and their positions. This method opens up new perspectives in stylometry: On the one hand, conjunctions as function words provide a large amount of data for statistical analyses, and on the other hand, they also carry (grammatical) meaning – about the relations between clauses (e.g. opposition, conditionality). By examining the relative frequency of each type, it is possible to reveal the most typical relations between clauses in a given text or corpus. In this way, the style of novels can be described at the level of the sentence, while also revealing the rhetorical-logical structure and epistemological attitude of the texts, which is not usually reflected in the reading process. This method also provides an opportunity to identify different stylistic traditions in literary history.

1. Introduction

This paper offers a stylistic analysis of the sentence structures of 150 canonical Hungarian novels published between 1832 and 2005. In doing so, it proposes a method for examining the frequency of different types of compound and complex sentences (in other terminology: clause linkage [Kabatek et al. 2010; Raible 2001], junction [Raible 1992] or clause complexes [Kugler 2020]), which can serve as a starting point for future studies on both literary history and the linguistic construction of literary texts. "Clause complexes profile multiple referential scenes and their relations, integrated into a single, complex structure. [...] [They] are not structures produced by creating and concatenating clauses, they are not derivable from their parts; rather, they can be interpreted in terms of construction types (schemas) and their instantiations." (Kugler 2020, p. 77) This characteristic highlights that compound-complex sentences can be examined not only from their syntactic structure, but also from a semiotic, pragmatic and discursive perspective. (Visapää et al. 2014b, p. 2–5) Such a "functional-topological framework" allows us not only to rely on a strict grammatical-syntactic taxonomy, but also to take into account the semantic-pragmatic dimension of relations. This is particularly important, since the traditional grammatical classification does not seem to be applicable in all cases – most contemporary theories also question the very division of subordination and coordination. This is shown both by cross-linguistic research (i.e. some of the grammatical types have no equivalent in other languages – Cristofaro 2014) and also within a language: for example Wolfgang Raible's analysis of the following example: "'On account of her illness, Joan remained at home.'" In this case, the relation that is

expressed is again very clear: causality. Nevertheless, it would make no sense to speak of ‘subordination.’” (Raible 2001, p. 595)¹

This is why the paper mainly draws on the grammatical *meanings* developed by clause linkages when analysing the different types from a quantitative point of view. Furthermore, this ‘grammatical meaning’ also carries information about the *rhetorical* structure of the text: Following Christian Matthiessen and Sandra A. Thompson, we can conclude that clause relations play an important role in the organization of discourses, and are in fact grammaticalized versions of the cohesive rhetorical relations between larger units in a text. (Matthiessen and Thompson 1988) “The cross-linguistic study of clause linkage markers and the observation that they tend to fall into clearly definable semantic-pragmatic sets has led linguists recently to characterize somewhat more fully than in the past the conceptual and rhetorical functions of many types of clause combining.” (Hopper and Closs Traugott 2003, p. 177) Considering all this, the hypothesis of the paper is that the relative frequency of clause relations in a given text (or group of texts) can help in determining which types of relations are the most characteristic of this text, which in turn sheds light on its underlying rhetorical and logical properties.² For example texts with a relatively high number of conditional relations clearly have a different epistemological attitude (i.e. they arrange elements of the outside reality differently) than texts that rely mainly on adversative coordinations. Similarly, the absence of a certain type of clause linkage can also reveal a great deal about a text. Such is the case of Miklós Mészöy’s early work from the corpus, where there is hardly any causative, inferential and explanatory relations between the clauses – stylistically this is how he is able to express the main philosophical insights of the French existentialist literature (first of all Albert Camus), and a chain of unmotivated actions (*action gratuite*). (See data in *Data availability* for details)

The research project that served as the basis for the study took a similar approach by calculating the mean and median sentence lengths of the same 150 canonical Hungarian novels (for details *Corpus*). The figures based on these measures show some trends, which already have certain implications regarding the major historical changes in the style of the Hungarian novel. [Figure 1](#) shows the mean and median sentence lengths (in terms of number of words) and the regression curve fitted to the data points, while [Figure 2](#) shows the same for 23 novels by famous Hungarian writer Mór Jókai published between 1846 and 1894. The negative slope in the 19th century, which can even be observed in Mór Jókai’s oeuvre, can be attributed not only to stylistic changes in the Hungarian literary tradition but also the growing role of the press, the widespread use of new writing tools, and reforms in how reading and writing were taught in schools. (Szemes 2020) This linear trend can be observed in the literature of several European languages (e.g. in Spanish – Calvo Tello 2023; and other languages: Schöch 2022).

In [Figure 1](#) the last third of the twentieth century is marked not so much by a definite trend as the co-existence of different types of prose: While extreme values are produced by authors associated with long sentences, some novels from this era employ distinctly

1. I will however use these terms throughout the paper for the sake of simplicity, but the focus will be on the individual relationship types.

2. The word ‘logic’ is used in a broad sense: It should not be understood as a term from formal logic but simply as a definite relationship between independent clauses. Therefore, it might be more accurate to use the expression ‘the diagrammatic character of sentences.’ Stjernfelt 2010

short sentences. Note, that the trendline is slightly overfitted due to the outliers (see the detection of the outliers in Appendix [Figure 11](#), and the overall trendline without them in [Figure 12](#)), but without outliers there is still a group of novels in the second half of the 20th century whose sentence lengths show a steady – linear – increase ([Figure 13](#) in Appendix). At the same time, these outliers are not in the corpus because of selection bias – the novels are in the center of the Hungarian canon from the 1970-1980s, the period called “prosa turn” (Szirák 2013, p. 504) with internationally recognized authors like the Nobel Prize winner Imre Kertész, Péter Nádas, who has been a contender for the prize for years, or the International Man Booker Prize winner László Krasznahorkai. The dispersion in the second half of the 20th century is illustrated in [Figure 3](#), where the texts are arranged chronologically and divided into groups of 30 texts, with the distribution of works consisting of “long”, “medium-length” and “short” sentences shown in the five resulting time frames. These three categories were created separately for each time frame based on average sentence lengths with the help of the k-means algorithm ($k=3$), following the method outlined in the 14th Pamphlet of the Stanford Literary Lab. That pamphlet illustrates the frequency of analepses and prolepses (flashbacks and flash-forwards) in movies; not by plotting the films along a single trend line but by dividing movies from each decade into three clusters based on whether they have “extreme”, “moderate”, or “conservative” values. (Kanatova et al. 2017) [Figure 3](#) indicates that from the 1970s onwards, works with long and short sentences start diverging more conspicuously than before, which suggests an increasing divide between coexisting prose styles. ([Figure 13](#) in the Appendix uses the same method and shows novels from the 20th century grouped into three parts based on chronology and clusters them in each group just into “high” and “low” categories without the outliers.) Furthermore this dispersion could be the reason why there is no statistically significant relationship between year of publication and sentence length in the whole data set according to the one-way analysis of variance (ANOVA) test ($p\text{-value} = 0.29$), just in the 19th century ($p\text{-value} < 0.001$) and in the whole dataset divided into three parts (1832-1899, 1900-1949, 1950-2005; $p\text{-value} < 0.001$)

Visualizations of mean sentence lengths can therefore capture literary and stylistic developments and help in distinguishing between short-sentence and long-sentence prose traditions. However, the similarities and differences of these traditions as well as the internal structure of the sentences should be examined more thoroughly, since ‘long sentence’ does not necessarily have the same meaning across different authors and periods. (Allison et al. 2013)

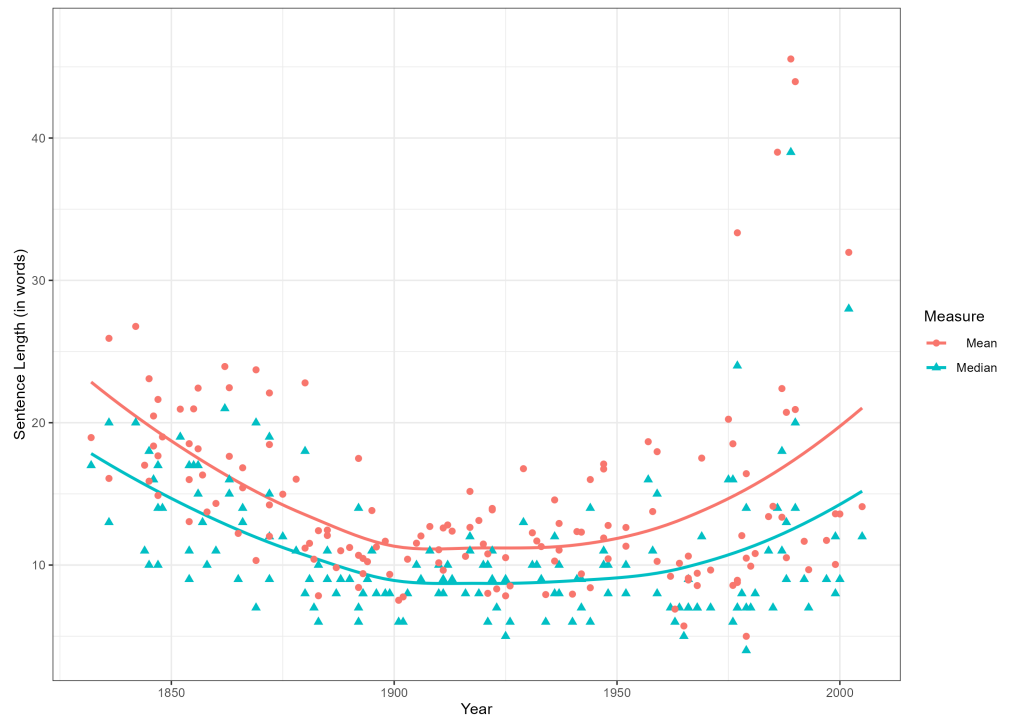


Figure 1: The mean and median of the sentence lengths of 150 Hungarian novels with loess trendline.

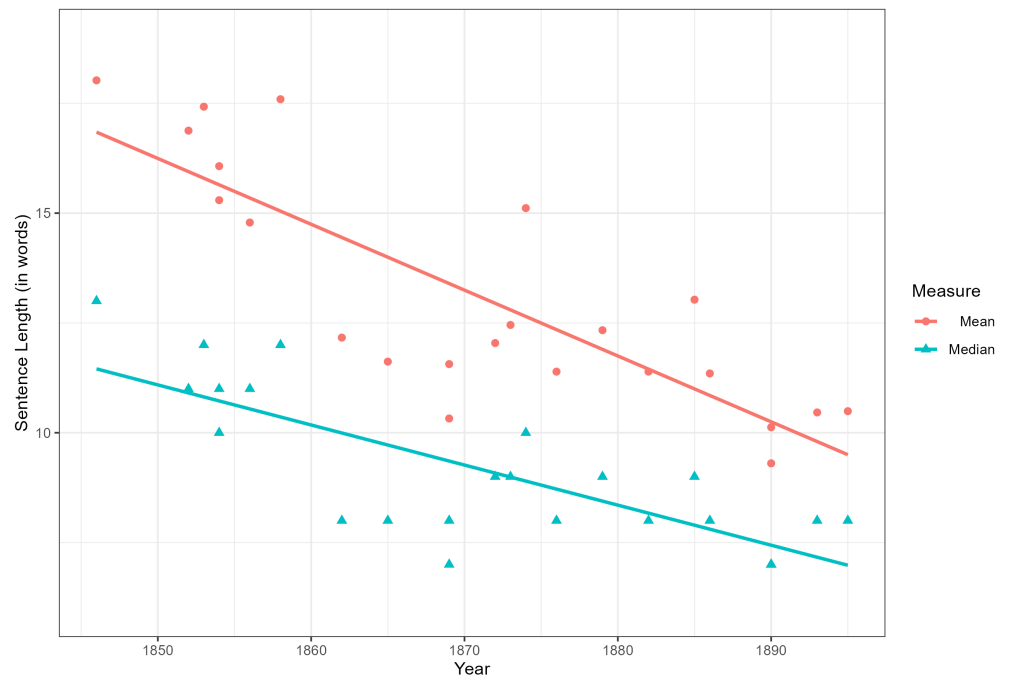


Figure 2: The mean and median of the sentence lengths of 23 novels by Mór Jókai with linear regression trendline. For mean $R^2 = 0.67$.

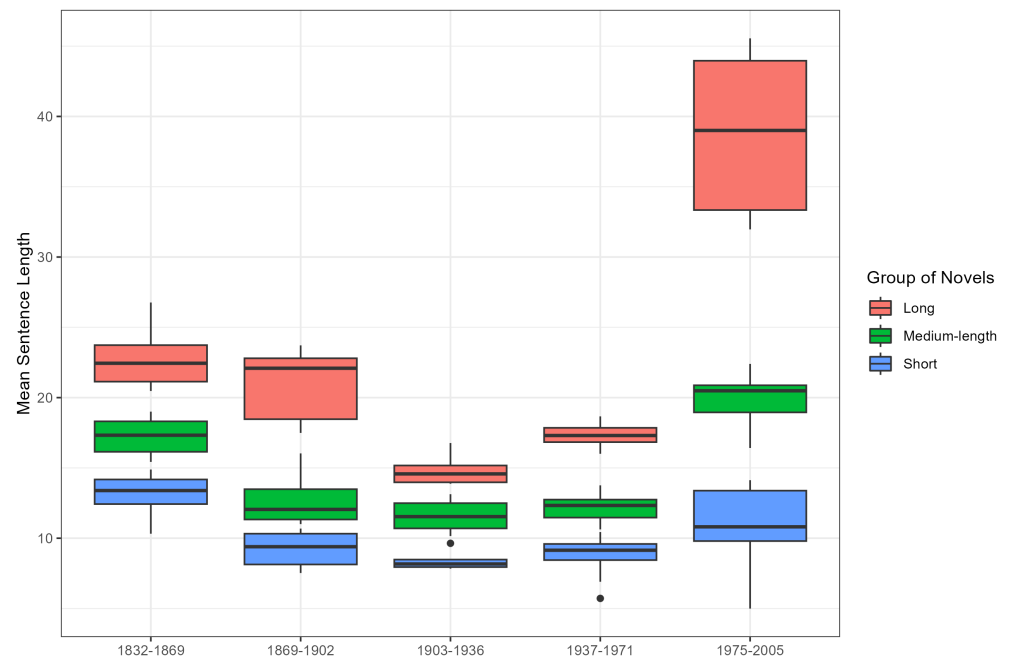


Figure 3: The distribution of novels with “long”, “medium-length”, and “short” sentences over five equivalent time frames, based on 150 novels.

2. Methods

In Hungarian orthography, clause boundaries are always marked by punctuation (commas, semicolons, colons, dashes), so identifying them is easier than in many other languages. Thus, combinations of punctuation marks and conjunction words or relative pronouns is enough to identify different clause relations, which can be detected with the help of basic regular expressions. Other scholars search for more complex grammatical features, which can also be used on unedited texts (e.g., transcriptions of colloquial language) and other languages. A more sophisticated method would be the practice of identifying conjunction words between finite verbs, since a clause prototypically consists of a finite verb as predicate and its elaboration. Although this might be a more general and nuanced approach, it is subject to three problems. Firstly, the *emagyar* NLP tool used for the present study (Váradi et al. 2018) cannot identify finite verbs with perfect accuracy, especially in texts from the first half of the 19th century. Secondly, not all conjunction words are placed between finite verbs; some of them can be found, for instance, at the beginning of a sentence. Thirdly, this approach does not take into account the cases in Hungarian where the clause is constructed not with a verbal but a nominal predicate (similar to the sentence: ‘That the house nice which tall.’ Dömötör et al. 2020). Another option would be to rely on the outcome of a dependency parser. However, their accuracy is rather low (Orosz et al. 2022) and, more importantly, they are based exclusively on syntactic categories (like “adverbial subordination”), while the research prefers to focus on the semantic and rhetoric dimensions of relations (*Introduction*).

For regular expressions, the following features are of particular importance: (1) the position of the conjunction words/relative pronouns, i.e., their possible distance from the punctuation mark; (2) the prototypical meaning of the polysemantic conjunction

words, and (3) whether the given conjunction word typically tends to link clauses or phrases. For example the conjunction *azonban* ('however') can be placed anywhere in the clause thanks to the flexibility of the Hungarian word order (*Okos gyerek volt, az élet értelmét megfejteni azonban nem tudta*; 'He was a clever child, he could not understand however the meaning of life') – while *de* ('but') only establishes adversative relation between clauses right after the comma (it prototypically develops such a relation between phrases anywhere else in the clause – e.g. 'He is tall but strong.')

We sought to answer these questions by manual analysis of 100-sentence samples from the corpus with the help of linguist annotators. If a conjunction word/relative pronoun created the same semantic type of clause relation at least 60 times out of 100 cases, it was classified under that type; otherwise, it was excluded from the study. For example, the conjunction word 'that' [*hogy* in Hungarian] was not included because multiple uses occur with similar frequency. (This has a significant impact on the analysis: In a sample of 1,000 compound-complex sentences with 2,502 clause combinations from the corpus, 12.3% of them are of the *that-type* [sum = 321]). The relatively high margin of error in the categorization is due to the ultimate goal of the present research, i.e. not to detect predefined grammatical categories in texts but to analyze the rhetorical-diagrammatical properties of novels, which requires as much data as possible. Thus, relations between phrases with a similar grammatical function and meaning to that of clause relations were not excluded from search results in all cases, since this allows better identification of the predominant types of relations. (However, it is also worth looking at these separately, as in the case of the conjunction *and*, which coordinates clauses more often in speech-related texts, while coordinates phrases more often in formal-written texts – Kytö and Smitterberg 2023.) The other extreme, that is, applying no limits and searching simply for conjunction words, is not efficient due to the problems caused by polysemantic words (e.g. *bár* means 'although' and 'bar' at the same time; it makes a difference whether the text expresses a concessive relationship or whether the characters are just thirsty), and by mixing semantically distinct grammatical structures, so the results would no longer reflect a clearly defined property of the novels. The application of this compromise between permissive and restrictive criteria is facilitated by Hungarian orthography, inasmuch as the relations between phrases are only marked with punctuation (and a pause in spoken language) if they have a grammatical meaning similar to that of clause relations. (e.g. *Hogy volt az embernek történelem előtti, azaz olyan korszaka, amelyről semmi, nemcsak írásban, de még szájról-szájra adott mondákban sem maradt fenn: az kétségen felül áll.* 'That there was a prehistoric age of man, that is, an age of which nothing has been preserved, not only in writing, but even in word of mouth: It is beyond doubt.' [Zsigmond Móricz, *Be Faithful Unto Death*])

During the research, conjunction words were classified into one of the following 12 categories (inspired by Seiler 1995 and Kortman 1997, adapted to Hungarian on the basis of Imrényi and Kugler 2018). After each category, the English translation of the most common conjunction word is given: 1. copulative ('and'), 2. adversative ('but'), 3. disjunctive ('or'), 4. inferential ('so', 'thus'), 5. explicatory (there is no strict equivalent in English grammar; 'namely', 'that is') 6. conditional ('if'), 7. concessive ('although'), 8. simile ('as', 'like'), 9. logical ('because'), 10. prototypical relative clause ('who', 'which'), 11. relative clause – space ('where'), 12. relative clause – time ('when').

A compound-complex sentence might fall into more than one category. For instance, the sentence ‘I see no contradiction in your response, and thus, if I think about it, you have to be right’ belongs to three categories, as it includes copulative, inferential, and conditional relations as well. Clause relations that are not marked by conjunction words or relative pronouns were disregarded in the study, for example: ‘There are no conjunction words in this sentence; [so] it was left out of the results.’ This draws attention to a crucial feature of the procedure: The research is not concerned with complex-compound sentences in general but only with those in which relations are elaborated grammatically. In the 1,000-sentence samples with 2,502 clause combinations, 35.3% of them belonged to the “not elaborated” category (sum = 883, while the number of linkages with conjunctions: 1,298, i.e. 51.9%) – so their exclusion is a strong limitation. But note they often include cases that do not primarily indicate a relation between clauses but a change of voice in the narration (e.g. “‘It’s hot”, said the snowman’), or an unmarked subordination of the *that*-type (‘I don’t know [that] when he’s coming.’) What is more important that the elaboration by a conjunction makes the connection between the clauses more accessible, i.e. it does not create the relation but “profiles it, makes the nature of the relation accessible and thus guides the interpretative process”. (Imrényi and Kugler 2018) By contrast, in the case of unmarked relations (often called juxtapositions, syntactically and semantically the less integrated type of connection, Raible 2001), multiple and subjective interpretations are possible, and the exact relationship between the clauses is less in the foreground for the reader. Moreover, conjunction words are particularly important for a quantitative analysis because, being function words, they are very common and thus provide enough statistical data to map the deep structures of a text (linking the method to previous researches in stylometry – Chung and Pennebaker 2007; Kestemont 2014; Rybicki and Eder 2011), yet, unlike several other function words, they have (grammatical) meaning, which makes the results about their distribution and frequency easy to interpret. The history of the style of the Hungarian novel as outlined with quantitative methods is based on this twofold nature of conjunction words and relative pronouns.

3. Corpus

The 1830s were chosen as the starting point for the corpus, as this is when the rise of the Hungarian novel in the modern sense occurred. The earliest novel in the corpus is András Fáy’s 1832 novel *The House of Belteky*, “traditionally regarded as the first Hungarian domestic novel of manners.” (Czigány 1984) The number and type of novels from the first half of the 19th century available in digital archives – mainly in Hungarian Electronic Library (Magyar Elektronikus Könyvtár – MEK) – had a major impact on text selection criteria and the size of the corpus. Since archives mainly store canonical novels from this era, only works that can be considered canonical are included in the corpus even from later periods. In any case, the few hundred novels available in databases would not give an overview of Hungarian fiction as a whole, but they are able to accurately represent the current literary canon. Canonicity was defined in two ways; if either criteria was applicable to a given text, it was considered canonical. First, the ELTE DH Novel Corpus (Bajzát et al. 2021)³ was consulted. This corpus, which

3. See: <https://regenykorpusz.elte-dh.hu/> (accessed Jan 17, 2023)

builds on MEK's database, currently contains 400 novels dating back to the 1920s and its first version has been created in the framework of the ELTeC (European Literary Text Collection) international COST-Action project.⁴ This project evaluates canonicity on the basis of publication history: Works are labelled as 'high canonicity' if they have had at least two new editions since 1979. Second, comprehensive studies of literary history were consulted, in particular in the case of novels published after the 1920s; works discussed in these studies were also considered canonical. It should also be noted that digitization and online availability can be seen as a form of canonicity: Texts that are included in online databases such as MEK and especially the prestigious *Digital Literary Academy* (*Digitális Irodalmi Akadémia, DIA*) are in some sense part of the literary canon. Indeed, only four novels from this study's final corpus are missing from these platforms. The corpus ends with Péter Nádas's 2005 novel *Parallel Stories*, excluding the last decades of contemporary Hungarian literature, as it would be difficult to find a criterion that would allow one to single out just a few canonical works from contemporary Hungarian literature.

The research, therefore, focuses on the history of the style of the Hungarian novel between 1832 and 2005. The corpus covering this period was created in two stages: First, 100 canonical novels by 58 authors were selected, with a minimum of 3 and a maximum of 8 novels from each decade, taking care not to over-represent any one period. To ensure proportional representation, a maximum of four novels per author was added to the collection, but preferably fewer; four novels were selected only if there was a significant time gap between them, which made it possible to examine whether the author's oeuvre followed a particular trend or whether the author had an artistic "fingerprint" unrelated to the period's trends. To examine this question more closely, a subcorpus consisting of 23 novels by Mór Jókai was created (e.g. [Figure 2](#)). In the second stage, 40 new authors and 50 new novels were added, bringing the total number of texts to 150 and the number of authors to 98, thus making the corpus more balanced both chronologically and in terms of authors, and providing an opportunity to double-check the previous measurements (i.e. whether the trends and patterns identified earlier could also be observed in the extended corpus). However, this comes at the price of including 19 new works that are not mentioned in major studies on literary history and are listed in the *ELTE DH Novel Corpus* and in the ELTeC as having 'low canonicity' – these novels are highlighted in the table of bibliographic data (*Data availability*). In other words, for the sake of a more detailed historical analysis, canonicity was de-emphasized when expanding the corpus. But, at the same time, since the corpus does not omit any author who is discussed in literary histories and whose works were published during the same time period as the 19 'low canonicity' novels, and since the vast majority of texts in 10 equal-sized timegroup is labelled as "high canonicity", the collection arguably remains representative of the current literary canon even after its expansion to 150 items.

On average, the corpus contains a work for every 1.15 years. To examine the distribution over time, the 173 years were divided into 10 equal parts so that the number of novels in each unit (17 years) could be counted and compared: 15 novels are in five groups, 14 in one, 13 in two, and 16 in two. Therefore, the difference between the periods with the highest and lowest values is only 3 novels. The period represented by the fewest works

4. See: <https://www.cost.eu/actions/CA16204/> (accessed Jan 17, 2023)

is the advent of the Hungarian novel (the period before the Hungarian Revolution and War of Independence of 1848-49): Far fewer works are available digitally from this era than from the second half of the century, which can be explained by the simple fact that fewer novels were written then. On average, 1.53 novels per author are included in the corpus: 64 authors are represented by a single text, 19 by two, 12 by three and 3 by four. Only 11 of the 98 authors are women, a disproportion that reflects imbalances of the Hungarian literary canon and institutional practices in the history of Hungarian literature.

For further bibliographic details of the novels see *Data availability*.

4. Results 1

After identification, we can calculate the relative frequencies of each type of clause relationship: Either relative to the length, i.e. the number of words in a novel; or relative to each other, i.e. proportionally. These values can be plotted in three ways: (1) focusing on historical changes of the relative frequencies along the timeline; (2) comparing the extent to which novels employ a certain type of relationship; and (3) concentrating on the proportions of the types within a single novel. In what follows, results for (1) and (3) will be reported.

Figure 4 shows the changes for prototypical relative clauses. The downward trend is caused both by the outliers of the 1840-50s and the fact that until the 1870s there is hardly any text in the lower regions of the graph. According to ANOVA there is a strong correlation between the frequency of relative clauses and the years of publication (p -value < 0.001), which means that the differences indicated by the trend can be considered statistically significant. This trend is present even without the outliers (for outlier detection see Appendix **Figure 11**, for the trend both with loess trendline and segmented linear regression see Appendix **Figure 14a**. **Figure 5** thus suggests a stylistic feature of the first half of the 19th century: Authors from this era tend to describe the characters appearing in the sentences in detail by using separate clauses (e.g., “The project, which was supported by the Foundation, is now finished.”)

Relative clauses are most frequently used in a rhythmic, rhetorical style of prose that was very influential in this period in Hungarian literature. This style is characterized by what might be called a ‘periodic sentences’; a compound sentences in which one part (in most cases from the main clause) is elaborated by several relative clauses outlining different scenarios (Herczeg 1981; for English context see: Carter and McRae 2005, p. 421). The following sentence (with the repetition of the subject) is taken from the novel *The Village Notary* from baron József Eötvös:

De ti, kiket nem bántottam soha életemben, s kik nyomorulttá tettetek, kik miatt nőm s gyermekim koldusbotra jutottak, kik belőlem gonosztevéőt csináltatok, kik kiűztetek az erdő vadjai közé, kik miatt e világon s az örökkévalóságban elkárhoztam...

‘you, who were the cause of my ruin! – you, who have caused my wife and children to beg their bread! – you, who made me a robber, who

hunted me, who compelled me to herd with the beasts of the forest!
(Eötvös 1850)

Another version of this construction is when different characters are elaborated by parallel subordinations in the same level of the sentence:

...az apai szív örömeiben dagadozva áldá a pogány író t s barátját, ki neki e könyvet ajándékozta, s mindazokat, kik elkészítésében részt vettek

'And old Esaias blessed the pagan author who wrote the book, and the college-chum who made him a present of it, and even the very printer who had produced it' (Eötvös 1850)⁵

Periodic sentences create a highly rhetorical language and make the text eloquent, but they can also be used for satirical-comical effect involving the accumulation of subordinations (especially in the passages criticising the Hungarian public conditions of the time.) This style of writing, despite its reliance on long sentences, is easy to understand due to its parallelisms, and has a strong rhetoric effect. In the Hungarian literary tradition it is significant until the last third of the 19th century – longer, than in Western European cultures (e.g. the decline its usage in England begins with William Wordsworth's and Samuel Taylor Coleridge's *Lyrical Ballads*, 1798 – Carter and McRae 2005, p. 420-22). It is because in Hungary, literacy education and the poetic tradition followed the ancient rhetorical model until the "Implementation of the Public Education Act" in 1868, which introduced compulsory education, and put the teaching of writing and reading on a new basis. Likewise, it is only in the second half of the century that a journalistic culture begins to emerge in which authors of texts based on shorter sentences and coordinations become successful (Mór Jókai played a leading role in this process – Figure 2).

According to the data, another trend can be observed in the mid-19th century, one that also employs complex sentences, where the relation between the clauses tends to be coordination rather than subordination. These clauses might appear in the text as separate sentences — connecting them with conjunctions reinforces the logical and/or causal relationship between them. Linking clauses this way creates a more loosely edited, irregular prose than using relative clauses in a periodic sentence, but it allows authors to depict a dynamic sequence of events and to elaborate on motivations and situations. This long-sentence style reappears in Hungarian fiction in the 1970s and in fact becomes even more characteristic of a group of writers (especially for the main figures of the "prosa turn"), as can be seen in the graph showing the changes in the frequency of inferential clauses (Figure 5, ANOVA p-value is at the significance threshold: 0.04.) The trend without outliers is shown in Appendix Figure 14b.

This figure and these results point to a similarity between certain novelists from different periods (and might even suggest cyclicity in the history of style). However, the differences between the two eras should also be noted. In the 19th century, inferential conjunctions usually introduce a sequence of events that logically follow each other and provide detail on a character's motivations. Making the successive nature of the events explicit is important because it clarifies the relation between complex structures that are

5. The English translation divides the original long sentences into several parts, so in selecting the examples I have focused on the English version to give a sense of the typical sentence structure of the original.

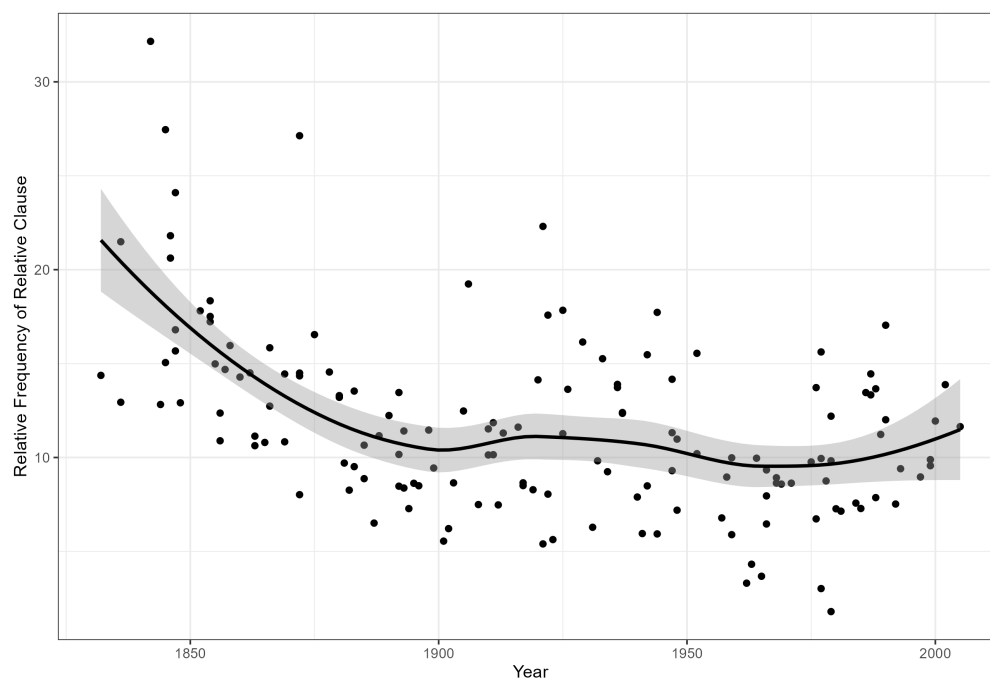


Figure 4: The changes of the prototypical relative clauses over time based on 150 novels with loess trendline, CI = 95%.

otherwise difficult to understand. Yet, as early as the 19th century, authors start using inference ironically and parodically; they do so by employing inferential coordinations or causal subordinations to make surprising connections between clauses that do not follow logically from each other. This can be observed in the following quotations from Ignác Nagy:

Az aratást tudniillik ma végezők be a dúsz alföld egyik gazdag birtokosánál a szegény emberek, mi természeteseb tehát, mint, hogy a vendégszerető fő táblabíró úr nemes szomszédait ünnepélyes lakomára hívó meg, mert hiszen illő, hogy az urak is kifáradjanak valamiben, miután a parasztok már derekasan megizzadtak.

‘The harvest was to be completed today by the poor people of the rich landowner, so what could be more natural than for the hospitable lord to invite his neighbours to a feast, for it is appropriate that the lords should also be tired after the peasants have sweated their hearts out.’ (Mosquitoes, my translation);

or:

Az árkokban rothadó víz és szemét igen ocsmány bűzt terjeszt és ez rendkívül hasznos, mert a faluról bejövő uraságokat arra figyelmezteti, hogy mielőbb siessenek ismét vissza egészséges falusi levegőjükhöz.

‘The water and garbage rotting in the ditches spreads a very foul stench, and this is very useful, because it warns men coming in from the village to hurry back to their healthy air.’ (Hungarian Secrets, my translation)

Focusing on the other end of the timeline, it should be noted that twentieth-century novels with long sentences exploit this subversive or ironic potential of inferences. These

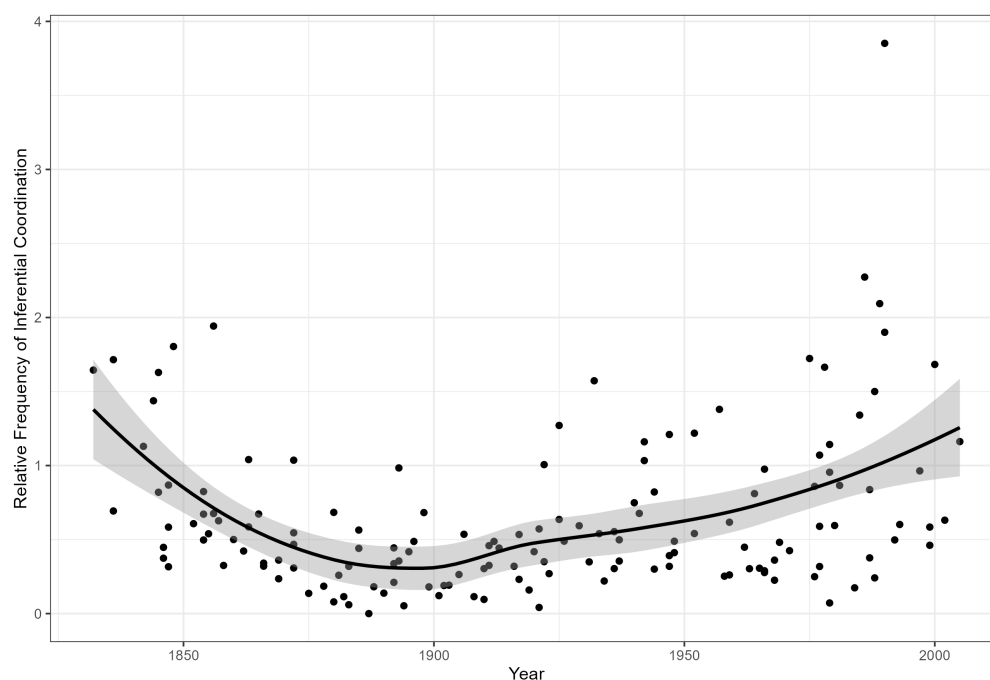


Figure 5: Changes in the frequency of inferential coordination over time based on 150 novels with loess trendline, CI = 95%.

works develop detailed, intricate relationships in complex sentences but draw attention to the artificiality or even absurdity of these relationships. The artificiality of the sentence structure may be due to the fact that writers of the era recognized that the language of narrative fiction was only one among many discourses and could only represent a world that had always already been interpreted a certain way. (Kulcsár Szabó 1994) Reacting to this recognition, novelists employ a high number of not only inferential coordinate clauses but also explicatory coordinate clauses – a stylistic trait not shared by nineteenth-century novels (see data in *Data availability*). The accumulation of explanations and inferences clarifies the logical structure of the sentence and the represented world, but at the same time, the sheer number of inferences and explanations also draws attention to the artificiality or inaccuracy of these relationships (since they can always be redescribed ever more accurately). This can be observed, for instance, Imre Kertész's, László Krasznahorkai's or Péter Nádas's novels in the corpus – which is closely related to the content of the texts. The works of Kertész for example describe events as the inevitable consequences of the functioning of the social order, and their aim is to explain this functioning as precisely as possible – whether it is the history of the Holocaust, Hungary of the 1950s or the Kádár era. The explanations show not only the natural consequence and cause of things, but also their uncanny absurdity:

Már egészen más dolog azonban – tették rögtön hozzá – az Arbeitslager, vagyis munkatábor: ott az élet könnyű, a viszonyok és az élelmezés, járt híre, hasonlíthatatlan, s ez természetes is, hisz ott a cél is más elvégre.

'An "Arbeitslager" or "work camp," on the other hand, it was immediately added, was something quite different: Life there was easy, the conditions and food, the rumors went, bore no comparison, which is natural enough as the aim, after all, is also different.' (Kertész 2006)

The same technique is applied in the *Kaddish for an Unborn Child* by Imre Kertész, where the English translation (similar to the quote above) uses disjunctive conjunctions and the term “more precisely” where the original employs explanatory relationships leading “to the point of absurdity”:

ez a kérdés te vagy, pontosabban én vagyok, de általad kérdésessé téve, még pontosabban (és ezzel nagyjából doktor Obláth is egyetértett): az én létezésem a te léted lehetőségeként szemlélve, vagyis én mint gyilkos, ha a pontosságot a végletekig, a képtelenig akarjuk fokozni, és némi önkínzással ez meg is engedhető, hiszen, hál' isten, késő, mindig is késő lesz már...

‘and you are that question; or to be more precise, I am, but an I rendered questionable by you; or to be even more precise (and Dr. Obláth, too, broadly agrees with this): My existence viewed as the potentiality of your being, or in other words, me as a murderer, if one wishes to take precision to the extreme, to the point of absurdity, and albeit at the cost of a certain amount of self-torment, since, thank God, it’s too late now...’ (Kertész 2004)

The same is true of László Krasznahorkai’s early prose, where the characteristics of the narrated world are also unfolded “from within”, from the characters point of view, which makes the otherwise absurd events reasonable and relatable. This internal perspective is an important difference from 19th century novels, which develop inferences and logical connections in a similar way and in similar numbers – but almost always from an external perspective. Thus the ironic effect is more clearly encoded in those narratives, in so far as they maintain an external perspective that can look at the existing conditions from the outside; in the case of Kertész and Krasznahorkai (and some of their contemporaries), by contrast, the irony of the inferences is only created by the reader, since the character voices do not reveal the bizarre nature of the operations described.

The trends discussed so far mainly concern long-sentence prose, with outliers at both ends of the timeline. In contrast, the relative frequency of similes reaches its peak in the first half of the 20th century (Figure 6). Although we cannot speak of a clear trend, it is apparent that several texts in this period show an exceptionally high value. Both the uncertainty of the trend and the group of outstanding texts from the same period are indicated by the fact that based on the ANOVA test, a correlation between similes and year of publication can only be found between three major periods (1832-1909, 1910-1949, 1950-2005; p-value < 0.001; without these yeargroups p-value = 0.84). But this historical tendency is confirmed, as it is also present without outliers – see Appendix Figure 14c both with loess trendline and segmented linear regression. An analysis of the corpus suggests that novels from this period develop explicit relations between clauses less frequently; instead, writers tend to describe situations by placing individual scenes or images side by side. An exception to this, however, is the use of similes, whose aim is precisely to link distant areas together, sometimes resulting highly poetic formulations (e.g. *Délfelé gyenge szél indul, forró, mint a gazdátlan büntudat*. ‘A light breeze comes up from the south, hot as uncontrollable remorse.’ – Miklós Mészöly, *Saulus* – my translation). The large number of these kind of relations may be related to the increasing emphasis from the 20th century on the representation of the inner lives of the characters and the ever more frequent use of the technique of free indirect speech. As

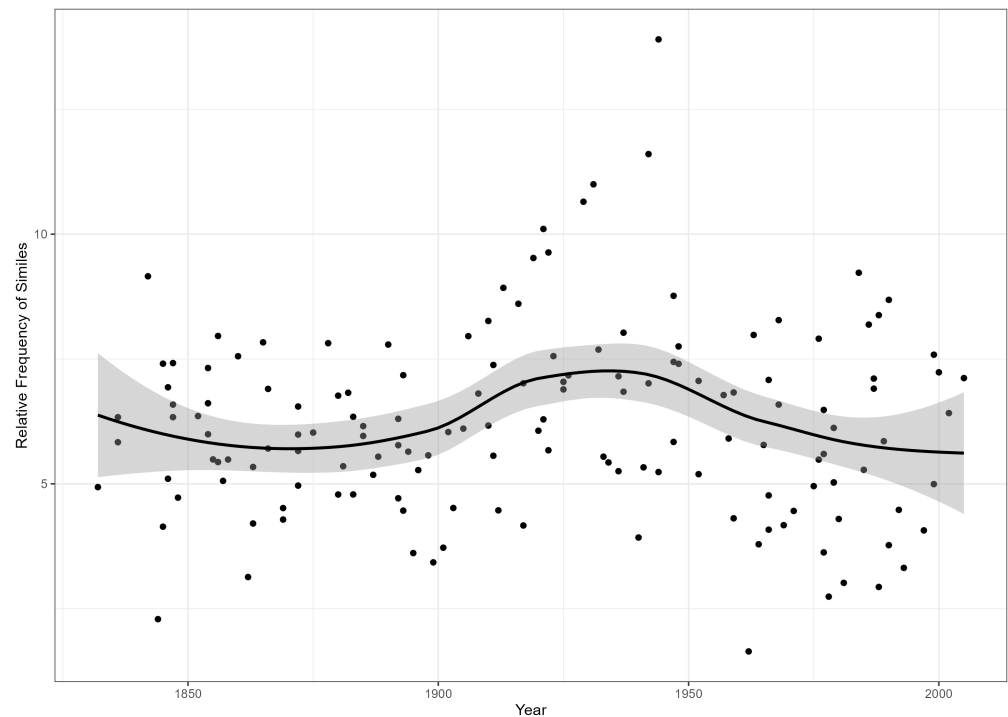


Figure 6: Changes of similes over time based on 150 novels with loess trendline, CI = 95%.

cognitive means of interpreting the world (Lakoff and Johnson 1980) and as expressions of subjective perception (“he felt as if he were enclosed in a glass enclosure”), similes are particularly suitable for incorporating these inner worlds into the narrative.

Moreover this graph identifies a new and chronologically distinct paradigm. While the use of relative clauses was prevalent in the first half of the 19th century and the use of explanatory and inferential relations abundant in the late 20th century (and, to some extent, in the first half of the 19th century), the first half of the 20th century is characterized by similes. These stylistic paradigms of sentence structure also hint at historical differences in the epistemological approach to the world. Relative clauses provide a more detailed description of a character in the main clause by representing them in a new scene, which fixes the meaning of the sentence. By contrast, constant clarification and explanation open up the sentence to new interpretations; this assumes a less fixed or semantically less anchored access to the world but offers a more subjective, internal point of view. Similes, the third paradigm, can provide new information by comparing two ideas, or simply pointing to their common features; in this case, the focus is not on characterization or elaboration but on creating an analogy (“You are beautiful like a rose”), articulating differences (“Your dad is not as strong as mine”), or incorporating the inner world of a character into the third person narration.

5. Results 2

Another way to visualize the results is to focus on the internal structure of the novels. In this case, one does not calculate relative frequencies in a text (relative to the number of words) but the percentage of the relation types in proportion to each other. Here the copulative coordinations are left out, partly because they provide the least relevant

information about the style of a text and partly because it is common for them to not be elaborated by conjunction words (e.g. juxtapositions like “I went to the restaurant, danced at the club, slept in the hotel.”) Thus, the results are only approximate and do not reflect the structures of the novels in their entirety – the figures only visualize the proportion of certain types of clause relations. For the sake of clarity, adversative and disjunctive coordinations; inferential and explicatory coordinations, and, finally, temporal and locative subordinations are put into joint groups in [Figure 7](#), since the grammatical function of these relations is quite similar. A more detailed figure could easily be made, but the resulting graph would be somewhat crowded. Similarly, one could also visualize the proportion of types within joint categories (e.g., the proportion of temporal and locative subordinations).

There are multiple reasons these graphs deserve just as much attention as the figures showing historical tendencies in the change of the frequencies relative to the number of words. First, while the previous figures examined the types separately, these graphs show the proportion of them at once. Since some types (e.g., concessive, explicatory, and inferential relations) are almost always less frequent than others (e.g., relative clauses), what the graphs reveal is not necessarily the most frequent type in a text but the extent of the difference between types. Secondly, the figures offer a better representation of the characteristics of the novels that either had high values in several categories or that did not have high values in any of them but whose internal structure is interesting for some reason (for example Iván Mándy’s novel *A Trafik* [*The Tabacconist*], where similes dominate a text that is otherwise poor in elaborated connections). Thirdly, these diagrams allow researchers to compare texts in a different way: Since every type of relation is shown in the same graph, all types can be taken into account in the comparison, and the difference between the proportions can be seen at a glance.

In this figure, one can easily see which works tend to use similar sentence structure. Moreover, the similarity between novels written by the same author (e.g., József Eötvös) is unmistakable. This raises the question of whether the quantitative analysis of clause relations can help not just in exploring stylistic trends and interpreting individual texts, but also in authorship attribution. (This hypothesis is supported by Grieve, who explains the difference between authors in terms of registers rather than idiolects; Grieve 2023.) In other words: To what extent can texts belonging to one author be distinguished from other authors on the basis of our data? Authorship researchers have shown on several occasions that there is a so-called “authorial fingerprint” beneath the thematic level of texts, which refers to the distribution of the most frequently – and therefore unconsciously – used words, mainly function words without specific meaning; and this distribution is approximately constant across texts of different genres written by a given author throughout his or her career (Baayen 2001; Burrows 2002; Rybicki and Eder 2011).⁶ This means that based on the relative frequency of the most frequent words one author is distinguishable from another (irrespective of their social and aesthetic backgrounds). The question is whether this rule holds true even when considering only conjunction words and those relative pronouns that form a connection between clauses.

6. “[W]e assume that in a language, there is a subset of (traceable) linguistic features dependent on an individual idiolect rather than shared by writers of the same epoch, genre, gender, etc. In a word, we believe that some features of a written text can betray the person who wrote it, despite his/ her aesthetic, social, or historical conditions.” (Eder 2011, p. 101)

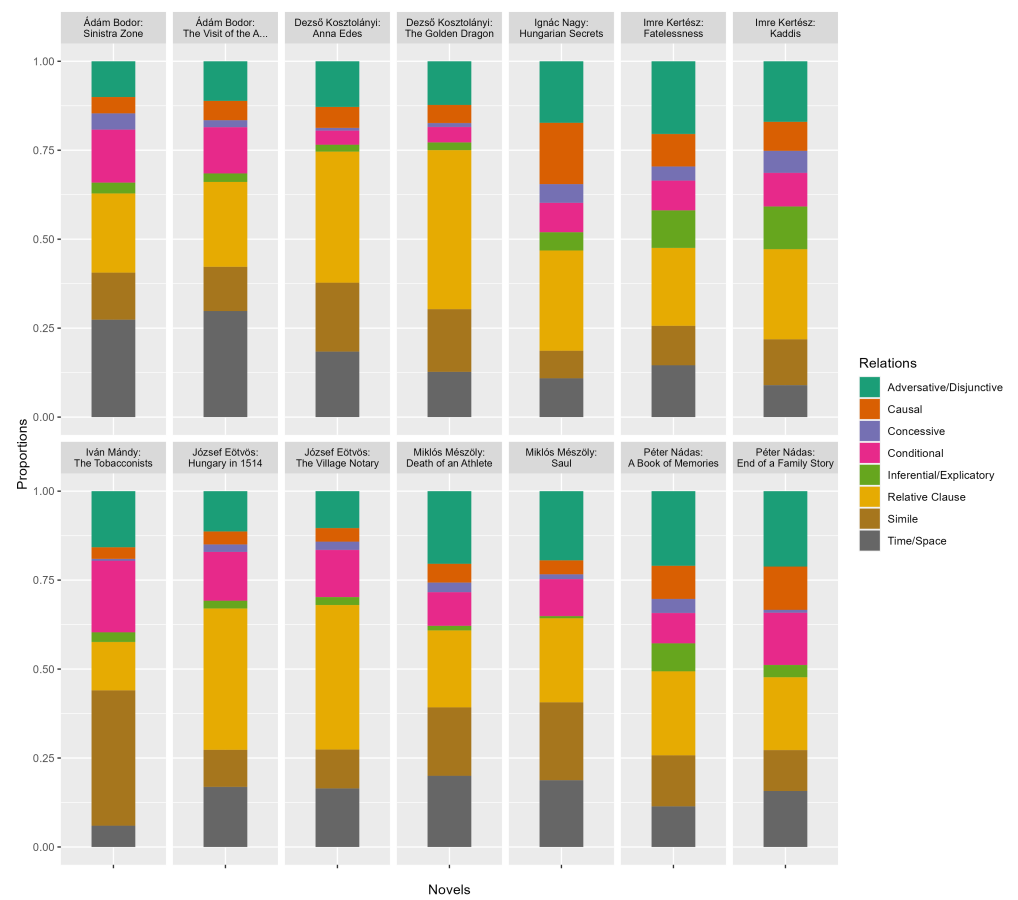


Figure 7: The internal structure of certain novels based on the categories under scrutiny.

To test this question, both the frequencies relative to the length and relative to each other were taken as a starting point: A novel is thus associated with several measures (12 for relative frequencies and 8 for proportions, respectively) which can be used to place the texts in a multidimensional space (i.e. a multidimensional coordinate system). Such a multidimensional space, however, cannot be represented or imagined – but it helps us to describe the similarity and dissimilarity of texts in two ways. The first possibility is not to plot the datapoints in this space, but simply calculate the distance between their positions according to metrics that work the same in two dimensions, then to group the texts based to their proximity (in this case, using Ward’s method), and finally to visualize these groups in the form of a dendrogram.

When grouping novels in this way, the performance of several distance metrics was compared. Cosine distance proved to be the most effective in terms of the two types of data (relative frequency and proportion), while Manhattan distance performed best when considering the aggregated data set. (However, there seems to be a consensus among scholars that cosine distance is the most reliable metric for authorship attribution. Evert et al. 2017) The latter phenomenon is illustrated by the dendrogram of [Figure 8](#), which shows 33 novels from the corpus, at least two of which have the same author (the same authorship is color-coded); this allows one to analyze the accuracy with which texts by the same author are placed on the same branch in the plot. In many cases, works by the same author are grouped together correctly, but there are also some misclassifications (Adjusted Rand Index, ARI = 0.54).⁷

The second possibility is to reduce the dimensions of the multidimensional space while preserving the spatial relationships of the data points as much as possible. The results of such a reduction based on principal component analysis (PCA) for relative frequencies are shown in [Figure 9](#). Here, the difference and similarity between texts is a function of the position and the percentage assigned to the axes (PC1 and PC2) – a value that shows the extent to which the distance on the axes plays a role in distinguishing data points. Thus, texts that are similar according to the selected criterion are positioned close to each other; while works with the same authorship are shown in the same color. In addition, each type of relationship is also marked as a loading in the figure, the direction of which shows how these types influence the location of the texts as data points.

The separation of novels by the same authors can be described as rather successful (i.e. we can support the hypothesis), even if not perfect: Dendrogram grouping does not work without errors, whereas in the PCA diagram these novels are in most cases in a similar position but cannot be clearly distinguished from other texts or groups of texts. This is due to the very characteristics under investigation. Namely, the use and the distribution of conjunction words and relative pronouns belongs to a *semi-conscious* level of the text. While the distribution of other function words (such as articles) is rather independent of the authorial decisions – which is why their frequency can bear the “fingerprints” of the author’s style – the frequency of the different sentence structures is not entirely

7. The relative effectiveness of clustering can be illustrated by comparing it with the results of traditional authorship attribution methods. When clustering the same novels with the stylo package based on the 100 most frequent words and cosine distance (without sampling), ARI = 0.83.

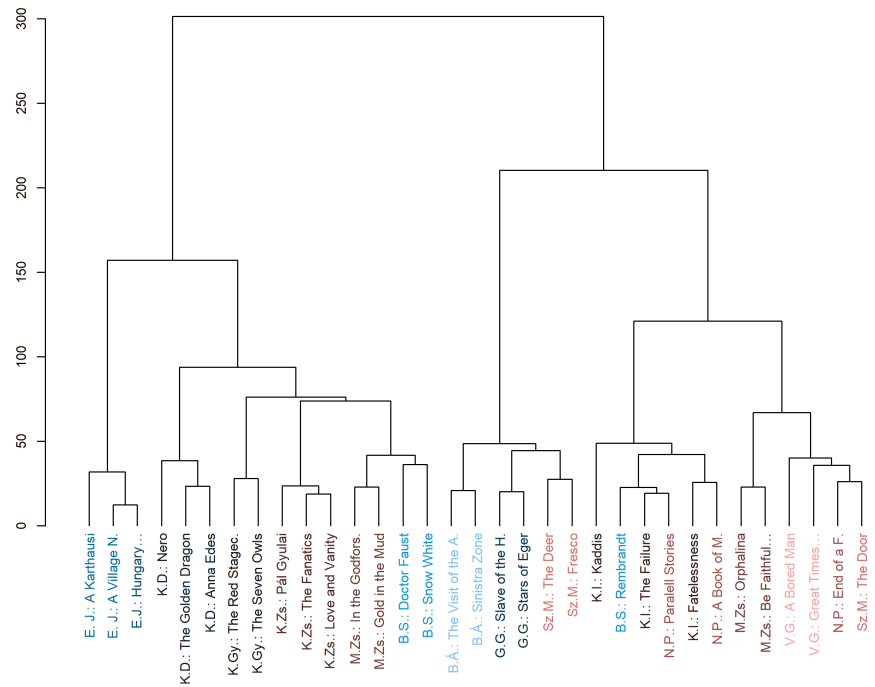


Figure 8: A clustering of 33 novels using Ward’s method based on the relative proportion and relative frequency of relation types – Manhattan distance. Authors and titles are abbreviated – for details see *Data availability*.

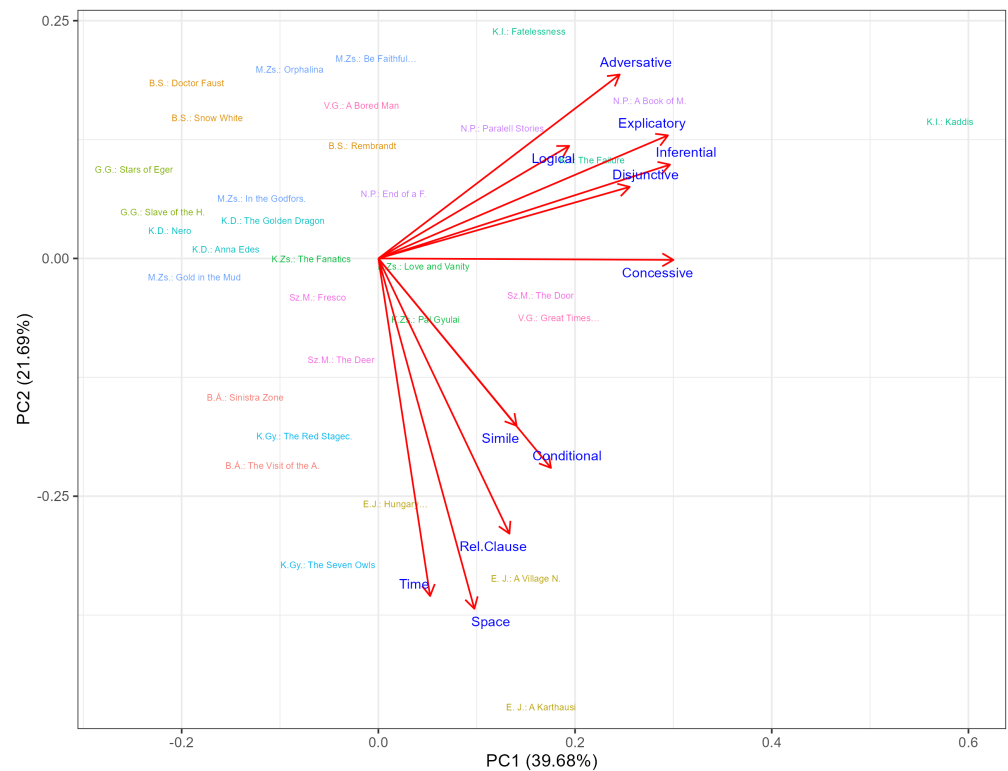


Figure 9: Principal component analysis by relative frequency of relation types, 33 novels.

independent of the author's individual considerations and aesthetic design.⁸ These words operate somewhere between the conscious and the unconscious levels, inasmuch as their use (unlike the use of content words) is not controlled in the creative process, but conscious authorial choices can influence their frequency. This also reflects that the question of which elements are under the author's control can be imagined along a continuum rather than along a conscious-unconscious dichotomy. Furthermore, this is why the experiment can be considered successful but without perfection, in that it is possible to group several texts by the same author, but at the same time it is not possible to group texts that are written in very different registers. Consequently, the analysis of the frequency of clause relations is above all not suited to answer questions of authorship attribution (or at least not by itself); its real use lies in identifying different stylistic traditions in the history of prose.

Figure 9 clearly shows that there are two distinct groups of relation types which could be characteristic for a text, that is, two directions can be distinguished with the help of the loadings: Clause linkages traditionally fall into the category of coordinations (adversative, disjunctive, inferential, explicatory and concessive) and causal subordinations (that elaborate similar logical relation than some of the coordinations, mainly inferential and explicatory) operate in one direction; whereas comparative, conditional, relative, temporal, and locative subordinations operate in the other. The same holds true when the investigation is carried out on all the 150 novels (Figure 10). These results suggest that three categories of complex and compound sentences exist in canonical Hungarian fiction – which are perhaps also indicative of three stylistic traditions, in so far as each category includes texts from different periods. The first category tends to develop logical relations between coordinate clauses, similarly to how logical value is attributed mainly to conjunctions in formal logic. The second employs a style that tends to give more information on the actors (whether human or non-human) or the circumstances of the depicted scenes; in these, it is the number of relative, conditional clauses and similes that are high. The third category includes novels that favor no type of clause relations; they prefer short sentences and mark relations between clauses less frequently. Needless to say, these styles should be seen as archetypes, and are rarely realized in pure form. Moreover, as we have seen earlier, these categories are also subject to change over time, which can contribute to the idea of continuing but ever evolving traditions.

8. The differences in the distribution of words and authorial intentionality are worth considering along a continuum rather than within sharp boundaries that divide our utterances into totally unconscious and conscious expressions. A more accurate description is to speak of expressions that are more or less easily/automatically recalled from short-term memory of the speaker - cf. Nini 2023.

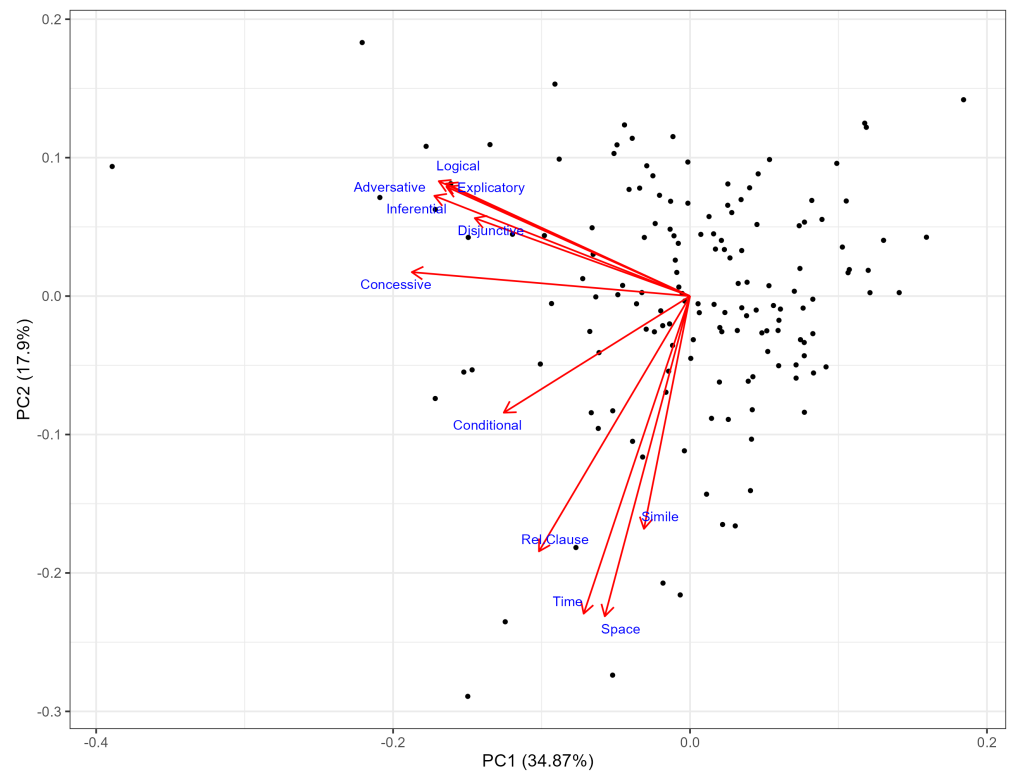


Figure 10: Principal component analysis by relative frequency of relation types, 150 novels.

6. Conclusion

An examination of the distribution of clause relations offers a better understanding of not only the linguistic structure of texts but also their diagrammatic, topological, and logical properties. Thus, the automatic identification of clause relations and the measurement of their frequency provides more than just a stylistic analysis: It can contribute to describing trends in literary history, interpreting individual novels, and distinguishing different traditions of prose style. The present study identified three traditions: A tradition that makes heavy use of subordinations, provides detailed descriptions of the elements in the main clause, and, thus, fixes the meaning of the sentence (mainly in the 19th century); a tradition that establishes logical relations between clauses, which keeps opening up the sentence to new interpretations (mainly in the second half of the 20th century); and a short-sentence tradition that relies chiefly on simple sentences, clause relations that are not elaborated, and similes (mainly in the first half of the 20th century). These conclusions are supported by various visualizations of the results; each type of visualization presents the values in a different layout to emphasize different aspects of the texts. Future directions for research include analyzing individual types in greater detail and breaking them down into subcategories, and complementing the research carried out so far by examining the relationships developed between sentences.

7. Data Availability

Data and code can be found here: https://github.com/SzemesBotond/sentence_structure.

8. Acknowledgements

The author would like to thank Péter Tamás and Ben Nagy for their help with the translation, and the members of the Computational Stylistics Group in Krakow and the Department of Digital Humanities in ELTE University, Budapest for their methodological and theoretical contributions.

9. Author Contributions

Botond Szemes: Conceptualization, Formal Analysis, Writing – original draft, Writing – review & editing

References

- Allison, Sarah, Marissa Gemma, Ryan Heuser, Franco Moretti, Amir Tevel, and Irena Yamboliev (2013). "Style at the Scale of the Sentence." In: *Stanford Literary Lab Pamphlets* 5. <https://litlab.stanford.edu/LiteraryLabPamphlet5.pdf> (visited on 11/06/2023).
- Baayen, Harald (2001). *Word Frequency Distributions*. Kluwer.
- Bajzát, Tímea, Botond Szemes, and Eszter Szláovich (2021). "Az ELTE DH Regénykorpusz és lehetőségei [The possibilities of the Novel Corpus from ELTE DH]". In: *Online térben az online térért. Networkshop 30: országos online konferencia. 2021. április 6-9, Eötvös Loránd Tudományegyetem [Hungarian] Online spaces for online spaceres. The 30th Networkshop conference*. Ed. by József Tick, Károly Kokas, and András Holl. HUNGARNET Egyesület. 10.31915/NWS.2021.7.
- Burrows, John (2002). "'Delta': A Measure of Stylistic Difference and a Guide to Likely Authorship." In: *Literary and Linguistic Computing* 17 (3), 267–287. 10.1093/l1c/17.3.267.
- Calvo Tello, José (2023). *The Novel in the Spanish Silver Age*. transcript. 10.1515/9783839459256.
- Carter, Ronald and John McRae (2005). *The Routledge History of Literature in English: Britain and Ireland*. Routledge.
- Chung, Cindy and James Pennebaker (2007). "The Psychological Functions of Function Words". In: *Social Communication*. Ed. by Klaus Fiedler. Taylor & Francis, 343–359. 10.4324/9780203837702.
- Cristofaro, Sonia (2014). "Is There Really a Syntactic Category of Subordination?" In: *Contexts of Subordination. Cognitive, Typological and Discourse Perspectives*. Ed. by Laura Visapää, Jyrki Kalliokoski, and Helena Sorva. John Benjamins Publishing Company, 73–93. 10.1075/pbns.249.03cri.
- Czigány, Lóránt (1984). *The Oxford History of Hungarian Literature : From the Earliest Times to the Present*. Clarendon Press. <https://mek.oszk.hu/02000/02042/html/index.html> (visited on 11/06/2023).
- Dömötör, Andreeq, Zijian Győző Yang, and Attila Novák (2020). "Much Ado About Nothing Identification of Zero Copulas in Hungarian Using an NMT Model". In: *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*.

- Ed. by Nicoletta Calzolari. European Language Resources Association. <https://www.aclweb.org/anthology/2020.lrec-1.591/> (visited on 11/06/2023).
- Eder, Maciej (2011). "Style-Markers in Authorship Attribution: A Cross-Language Study of the Authorial Fingerprint." In: *Studies in Polish Linguistic* 6, 99–114.
- Eötvös, József (1850). *The Village Notary*. Trans. by Otto Wenckstern. Longman, Brown, Green and Longmans.
- Evert, Stefan, Thomas Proisl, Fotis Jannidis, Isabella Reger, Steffen Pielström, Christof Schöch, and Thorsten Vitt (2017). "Understanding and Explaining Delta Measures for Authorship Attribution". In: *Digital Scholarship in the Humanities* 32 (2), ii4–ii16. [10.1093/l1c/fqx023](https://doi.org/10.1093/l1c/fqx023).
- Grieve, Jack (2023). "Register Variation Explains Stylometric Authorship Analysis". In: *Corpus Linguistics and Linguistic Theory* 19 (1), 47–77. [10.1515/cllt-2022-0040](https://doi.org/10.1515/cllt-2022-0040).
- Herczeg, Gyula (1981). *A XIX. századi magyar próza stílusformái* [*Stylistic Forms of 19th Century Hungarian Prose*]. Akadémiai Kiadó.
- Hopper, Paul J. and Elizabeth Closs Traugott (2003). *Grammaticalization*. Cambridge UP. [10.1017/CB09781139165525](https://doi.org/10.1017/CB09781139165525).
- Imrényi, András and Nóra Kugler (2018). "Mondattan [Hungarian] Grammar of Sentences". In: *Nyelvtan*. Ed. by Gábor Tolcsvai Nagy. Osiris.
- Kabatek, Johannes, Philipp Obrist, and Valentina Vincis (2010). "Clause Linkage Techniques as a Symptom of Discourse Traditions: Methodological Issues and Evidence from Romance Languages". In: *Syntactic Variation and Genre*. Ed. by Heidrun Dorgeloh and Anja Wanner. De Gruyter Mouton, 247–276. [10.1515/9783110226485.2.247](https://doi.org/10.1515/9783110226485.2.247).
- Kanatova, Maria, Alexandra Milyakina, Tatyana Pilipovec, Artjom Shelya, Oleg Sobchuk, and Peeter Tinitis (2017). "Broken Time, Continued Evolution: Anachronies in Contemporary Films." In: *Stanford Literary Lab Pamphlets* 14. <https://litlab.stanford.edu/projects/broken-time/> (visited on 11/06/2023).
- Kertész, Imre (2004). *Kaddis for an Unborn Child*. Trans. by Tim Wilkinson. Random House.
- (2006). *Fatelessness*. Trans. by Tim Wilkinson. Vintage.
- Kestemont, Mike (2014). "Function Words in Authorship Attribution. From Black Magic to Theory?" In: *Proceedings of the 3rd Workshop on Computational Linguistics for Literature (CLFL)*. Ed. by Anna Feldman, Anna Kazantseva, and Stan Szpakowicz. Association for Computational Linguistics. [10.3115/v1/W14-0908](https://doi.org/10.3115/v1/W14-0908).
- Kortman, Bernd (1997). *Adverbial Subordination. A Typology and History of Adverbial Subordinators Based on European Languages. (Empirical Approaches to Language Typology, 18.)* De Gruyter Mouton. [10.1515/9783110812428](https://doi.org/10.1515/9783110812428).
- Kugler, Nóra (2020). "Contextualizing Clauses." In: *Studia Linguistica Hungarica* 32, 76–90.
- Kulcsár Szabó, Ernő (1994). *A magyar irodalom története 1945–1991* [*History of Hungarian Novel 1945–1991*]. Argumentum.
- Kytö, Merja and Erik Smitterberg (2023). "Clausal and Phrasal Coordination in Recent American English." In: *Corpus Linguistics and Linguistic Theory* 19 (1), 23–46. [10.1515/cllt-2022-0035](https://doi.org/10.1515/cllt-2022-0035).
- Lakoff, George and Mark Johnson (1980). *Metaphors We Live By*. University of Chicago Press.
- Matthiessen, Christian and Sandra A. Thompson (1988). "The Structure of Discourse and 'Subordination'". In: *Clause Combining, in Grammar and Discourse*. Ed. by John

- Haiman and Sandra A. Thompson. John Benjamins Publishing House, 275–331. [10.1075/tsl.18.12mat](#).
- Nini, Andrea (2023). *A Theory of Linguistic Individuality for Authorship Analysis*. Cambridge University Press.
- Orosz, György, Zsolt Szántó, Péter Berkecz, Gergő Szabó, and Richárd Farkas (2022). “HuSpaCy: An Industrial-Strength Hungarian Natural Language Processing Toolkit”. In: *XVIII. Magyar Számítógépes Nyelvészeti Konferencia*. [10.48550/arXiv.2201.01956](#).
- Raible, Wolfgang (1992). *Junktion. Eine Dimension der Sprache und ihre Realisierungsformen zwischen Aggregation und Integration*. Winter.
- (2001). “Linking Clauses”. In: *Language Typology and Language Universals*. Ed. by Armin Burkhardt, Hugo Steger, and Herbert Ernst Wiegand. De Gruyter Mouton, 590–617.
- Rybicki, Jan and Maciej Eder (2011). “Deeper Delta across Genres and Languages: Do We Really Need the Most Frequent Words?” In: *Literary and Linguistic Computing* 26.3, 315–321. [10.1093/lc/fqr031](#).
- Schöch, Christof (2022). “Sentence Length across ELTeC Collections and Gutenberg Fiction”. In: *Distant Reading Closing Conference, April 21-22, 2022*. <https://christof.s.github.io/krakow22/> (visited on 11/06/2023).
- Seiler, Hansjakob (1995). “Junktion – Zu Wolfgang Raibles gleichnamigem Buch”. In: *Vox Romanica* 54, 12–21.
- Stjernfelt, Frederik (2010). “The Extension of Peircean Diagram Category: Charting the Implications of a Diagrammatical Revolution in Semiotic”. In: *Studies in Diagrammatics and Diagram Praxis*. Ed. by Olga Pombo and Alexander Gerner. College, 57–73.
- Szemes, Botond. (2020). “Mondathosszúság és irodalomtörténet [Sentence Length and Literary History].” In: *Literatura* 46 (3), 335–367.
- Szirák, Péter (2013). “Im Sog des Schrifttextes. Der literalistic turn in der ungarischen Nachmoderne ab 1960/1970”. In: *Geschichte der ungarischen Literatur*. Ed. by Ernő Kulcsár Szabó. De Gruyter, 502–548. [10.1515/9783110241105.502](#).
- Várad, Tamás, Eszter Simon, Bálint Sass, Iván Mittelholcz, Attila Novák, Balázs Indig, Richárd Farkas, and Veronika Vincze (2018). “E-magyar – A Digital Language Processing System”. In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Ed. by Nicoletta Calzolari, Khalid Choukri, Christopher Cieri, Thierry Declerck, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga. European Language Resources Association.
- Visapää, Laura, Jyrki Kalliokoski, and Helena Sorva, eds. (2014a). *Contexts of Subordination. Cognitive, Typological and Discourse Perspectives*. John Benjamins Publishing Company. [10.1075/pbns.249](#).
- (2014b). “Introduction”. In: *Contexts of Subordination. Cognitive, Typological and Discourse Perspectives*. Ed. by Laura Visapää, Jyrki Kalliokoski, and Helena Sorva. John Benjamins Publishing Company, 1–16. [10.1075/pbns.249.01her](#).

A. Appendix

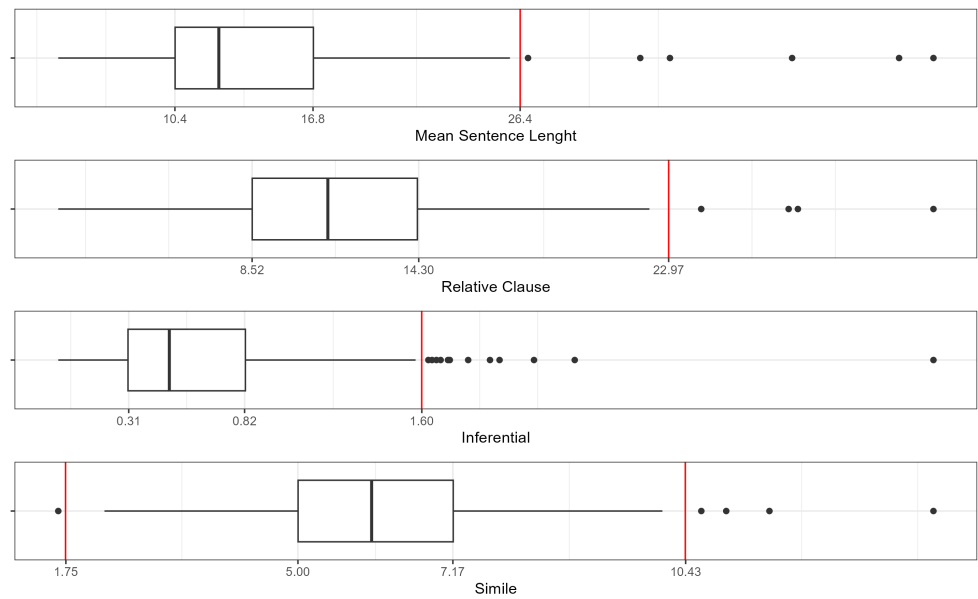


Figure 11: Outlier detection for mean sentence length, prototypical relative clause, inferential coordination and similes based on $Q3 + 1.5 * IQR$, and $Q1 - 1.5 * IQR$. The threshold values can be used to identify which novels are considered outliers. For details see *Data availability*

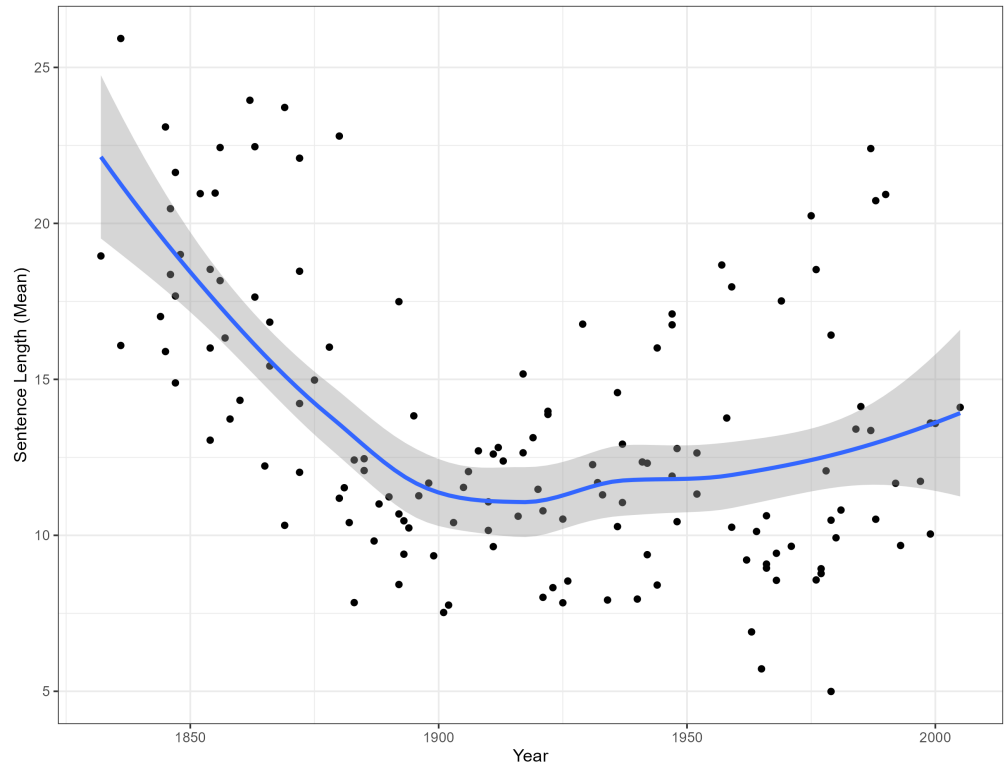


Figure 12: Trend in the changes of mean sentence length without the outliers (1832-2005).

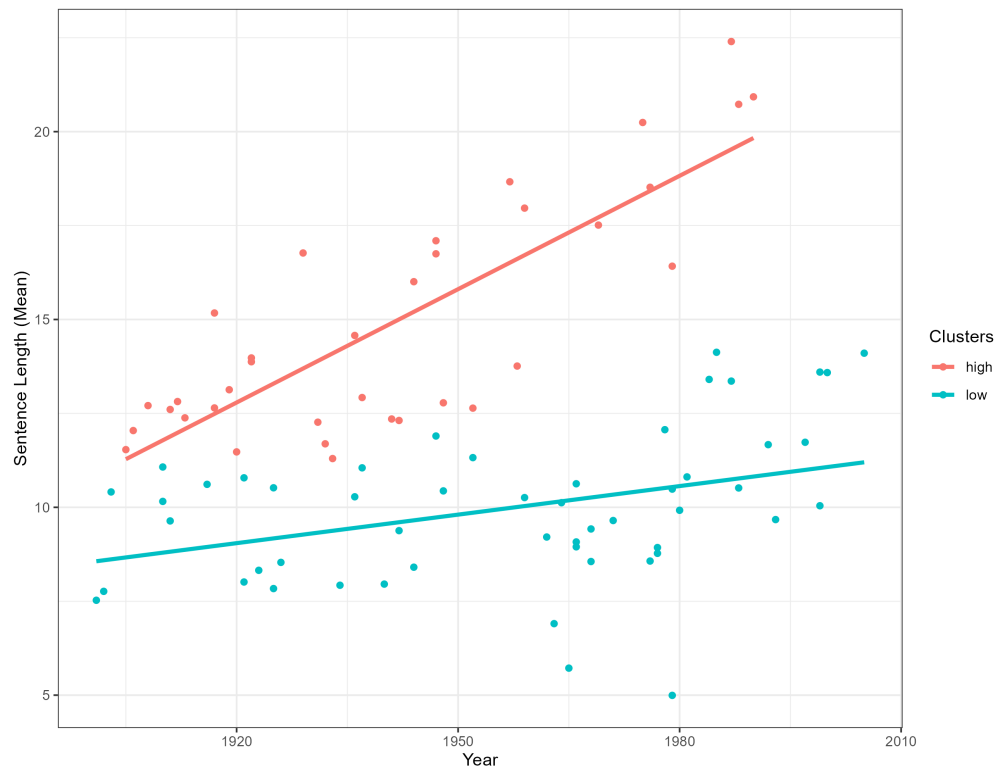
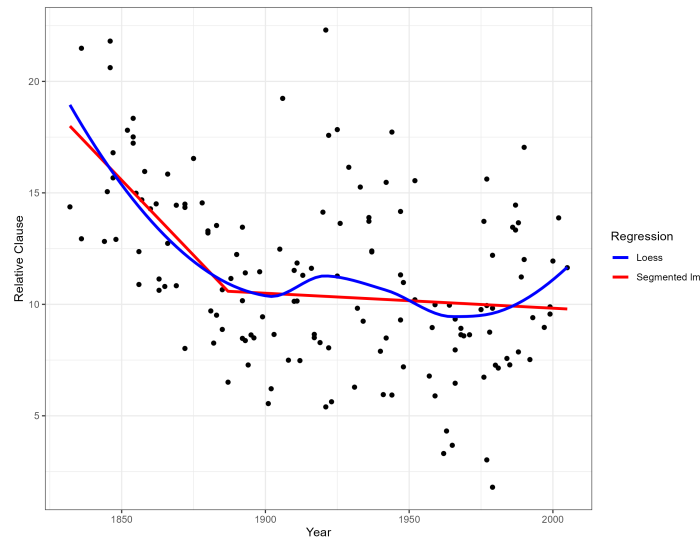
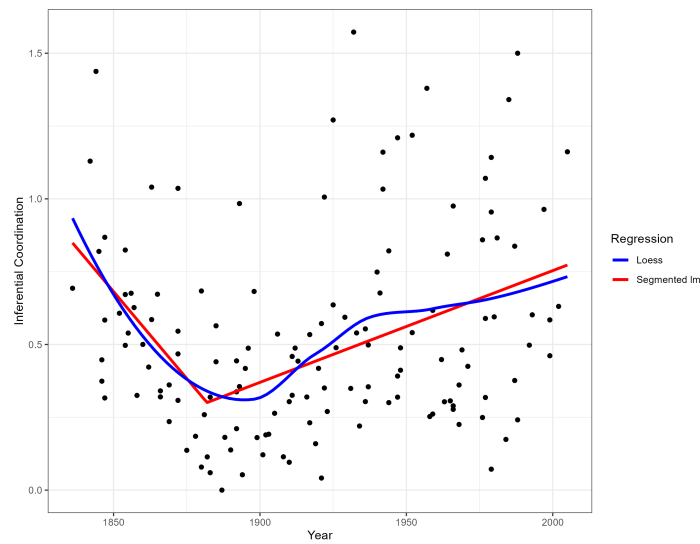


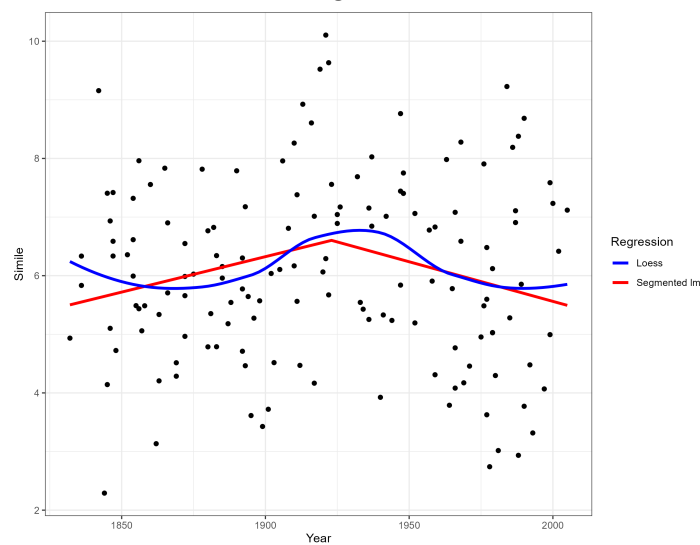
Figure 13: Trend in the changes of mean sentence length without the outliers (1900-2005), after clustering the data as described for Figure 3.



(a) Prototypical relative clause; for segmented lm $R^2 = 0.24$.



(b) Inferential coordination, for segmented lm $R^2 = 0.14$



(c) Simile, for segmented lm $R^2 = 0.02$.

Figure 14: Trends in a given clause relation without outliers based on loess trendline and segmented linear regression. Segmentation was done automatically using the *segmented* R package.