# D5.2

# CASE STUDIES IN DATA PREPARATION AND SHARING

Authors: MICHAŁ MRUGALSKI, ANNIKA BLIETZ, INGO BÖRNER, ELISABETH BAUER, VERA CHARVAT, MATEJ ĎURČO, SABINE LASZAKOVITS, STEFAN RESCH

May 31, 2023

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101004984

Project Acronym: CLS INFRA
Project Full Title: Computational Literary Studies Infrastructure
Grant Agreement No.: 101004984

## Deliverable/Document Information

Deliverable No.: D5.2
Deliverable Title: CASE STUDIES IN DATA PREPARATION AND SHARING
Authors: MICHAŁ MRUGALSKI, ANNIKA BLIETZ, INGO BÖRNER, ELISABETH BAUER, VERA CHARVAT, MATEJ ĎURČO, SABINE LASZAKOVITS, STEFAN RESCH
Dissemination Level: PUBLIC

## Document History

| Version/Date | Changes/Approval | Author/Approved by |
|---|---|---|

Table of Contents

## 1. Introduction

This deliverable presents three case studies involving digitisation and transformation processes; the studies are presented in order of the complexity of the research question, which is reflected in the difficulty of the corpus compilation task. Transformation processes seem to be inevitable in each case, but paradoxically the urgency of digitisation diminishes as the complexity of a task increases, so that in the simplest case of the Slovak novel one of the main obstacles – besides the lack of metadata or the poor technical quality of the available material – is the small number of texts available online to choose from.

But one thing after another. The case studies described in this deliverable are:

1. Creation of an ELTeC affine corpus of the Slovak novel (chapter 2)
2. Finding the haiku across multilingual corpora (chapter 3)
3. Measuring entropy and surprisal in the prose of the Tsarist Empire Devoted to Terrorism (Russian and Polish Texts) (chapter 4)

The first two case studies have already served as reference cases for the data landscape review (CLS INFRA Deliverable 5.1). This extended version, which conveys the experience of six months of research and is enriched by the third case study, highlights specific aspects of the multidimensional landscape of literary text collections. In Deliverable 5.1, they were merely illustrations and concretisations of general points; now they are the focus of attention. The third case has been designed with the most complex research questions in mind, to go even further in exploring what is available and what is possible in the digital humanities today.

The Slovak novel is a local form associated with a small Central European culture. The history of this part of Europe still affects the representativeness and balance of the potential corpus in a way that is hard to imagine for the inhabitants of cultural metropolises. The first use case therefore focuses on the problems of accessibility and discoverability of resources under conditions of resource scarcity. In the long run, this use case also raises the question of the availability of NLP tools for smaller languages, which is related to the question of their universality. In contrast to the Slovak novel, the haiku serves as an example of an international form typical of high modernism in the twentieth century literary cultures of the West. With the haiku case we enter the world of cultural and linguistic incommensurability, which overlaps with the Babel-like conditions of metadata and corpus architectures. The problem is no longer the scarcity of resources, but rather the precision of browsing tools: how to find exactly what we are looking for in large and diverse corpora. Finally, the third case study, devoted to the novel of terrorism in Polish and Russian literature (as variants of the literature of the tsarist empire), relates the issues of findability and accessibility in a multilingual milieu of less popular languages to refined research questions that could really contribute to the development of narratology and literary studies at large. We observe how an actual research hypothesis impacts the process of corpus compilation.

In order to stay in touch with current developments in the computational literary studies and computational linguistics, we have drawn on the experience of two research projects in the first and third use cases: The case of the Slovak novel corresponds to a research project conducted at the Slovak Academy of Sciences and the University of Nitra. In the second chapter, I quote our exchanges with the project leader, Dr Marek Debnár from the University of Nitra. The third use case revolves around the issues of surprisal and entropy as they are reflected in style (register). Most of the computational aspects of the projects rely on the results of the Information

Density and Linguistic Encoding (IDeaL) research cluster at the University of Saarland in Saarbrücken.

While the first case reveals mostly hidden political and economic underpinnings of the digital humanities field that could easily be addressed with additional resources, the second case presents a technical and conceptual solution currently by the development the CLS Infra project to the problems faced by researchers trying to explore today's data landscape, namely the Corpus Exploration Platform (CEP). While the second case study highlights the platform's most general and essential features, the third case points to directions for future refinement of both resource cataloguing methods and natural language processing tools.

## 2. The Slovak Novel

In this section we will first describe the general constructive assumptions behind ELTeC and then the ways in which these assumptions can be seen as problematic in relation to the Slovak novel.

### 2.1. Background: ELTeC and the Issues of Corpus Architecture

The first case study concerns the compilation of an ELTeC[1]-affine corpus of the Slovak novel. ELTeC stands for the European Literary Text Collection. It is one of the main results of the COST Action[2] "Distant Reading for European Literary History" (CA16204), which ran from 2017 to 2022. Work on the corpus highlights the fragmentary nature of today's data landscape, both in terms of the technical accessibility of resources that have fallen outside the scope of the world literary canon, and the cultural contingencies that undermine our notions of eligibility, representativeness, and balance.

ELTeC is a collection of literary text corpora in a number of European languages. The creation of the collection was seen as an essential prerequisite for the development and evaluation of multilingual tools and methods for the analysis of literary texts, including but not limited to authorship attribution, topic modelling, character network analysis or stylistic analysis. All ELTeC corpora are encoded in TEI-XML according to one of the following ELTeC schemas.[3]

In order to prevent excessive eccentricity that could jeopardize research based on counting similar aspects, the novels were chosen from the major literary genres on the basis of their size, availability, and communicative quality of their language. The novel is currently defined as "at least 10,000 words of fictional prose narrative."[4] The collection includes items from 1840 to 1920. The copyright restrictions (*terminus ante quem*) and the availability of high-quality full texts (*terminus post quem*) are what cause the chronological limits: When used on older printed material, OCR techniques frequently fail.

[1] https://www.distant-reading.net/eltec/; https://distantreading.github.io/ELTeC/; https://distantreading.github.io/; https://github.com/COST-ELTeC/ELTeC

[2] COST Actions are bottom-up networks that bring together researchers and innovators to investigate a topic of their choice over a four-year period. COST Actions typically involve researchers from academia, SMEs, public institutions, and other relevant organisations or interested parties.

[3] https://distantreading.github.io/Schema/eltec-0.html; https://distantreading.github.io/Schema/eltec-1.html; https://distantreading.github.io/Schema/eltec-2.html

[4] https://distantreading.github.io/sampling_proposal.html

ELTeC has three components: ELTeC core, ELTeC plus and ELTeC extensions; in the following we only refer to ELTeC core.[5] Each of the ELTeC core corpora contains a selection of 100 texts and is balanced to be as comparable with other corpora of the collection as feasible in size and composition. The selection criteria for the writings include length, popularity, gender, and the dispersion of the authors' ages.[6] There are four equal groupings of twenty years each that cover the period of novels written between 1840 and 1920. A work's popularity is determined by how many times it was reprinted between 1970 and 2010, as recorded in WorldCat or another appropriate national library catalog. In contrast to male and mixed-gender authors, a female author shall appear on no fewer than 10% and no more than 50% of titles. No more than three titles per author are permitted, and precise three books by no fewer than nine and no more than eleven authors are permitted. Only one novel is contributed by each of the other authors. Finally, at least 20% of documents should fall into each of the three length categories: short (10k–50k word tokens), medium (50k–100k word tokens), and large (>100k word tokens).

Each corpus is supposed to follow the same criteria for corpus composition, but some deviations were pre-programmed from the beginning. Because of this, the E5C score evaluates how closely the corpus criteria were followed.

For historical reasons,[7] the Slovak novel is doomed to perform poorly in this respect. This sheds some light on the ELTeC criteria and the tacit assumptions of the CLS in general regarding eligibility, representativeness, and balance as sought-for qualities of corpora – assumptions that these criteria reflect. The case of the Slovak corpus highlights the fragmented nature of today's data landscape in terms of both technical and intellectual accessibility (including inadequate metadata) of resources that are not produced by global players in the field of culture. This case study supplements the reflection on the cultural and political implications of the ELTeC criteria that the coprus creators provided in the case of French, Romanian, Portuguese, and Slovenian corpora: “In each case, the literary tradition, the history of the language, the current state of digitization of cultural heritage, the resources available locally, and the scholars’ training level with regard to digitization and corpus building have been vastly different” (Schöch et al. 2021). The researchers conclude:

> Regarding corpus building, it has become increasingly clear that equally strict adherence to all corpus-composition principles in all collections is almost impossible to achieve. On the one hand, the corpus-composition criteria have turned out to be very demanding relative to the actual production in many literary traditions during the period 1840–1920. On the other hand, there remains a tension between representativeness, composition principles, and comparability of collections. [...] it has become clear that the corpus-composition criteria are a double-edged sword. On the one hand, they provide a definite incentive to avoid repeating the biases encountered in many corpus-building endeavours—we have made sure, for instance, that we include a certain minimum proportion of novels by female authors and of non-canonical novels, and the fact that we have undertaken this effort makes ELTeC unique. [...] On the other hand, respecting the corpus-composition criteria is clearly more challenging for some languages than for others. Strict criteria favour better-resourced languages (such as English, German, and French), where finding 100 novels respecting all criteria is possible; it disadvantages lesser-resourced languages, where more challenges present themselves to reach the full size of 100 novels while including, for example, a

[5] ELTeC plus contains smaller collections of texts covering the same time period as ELTeC core, but which do not meet the balance criteria defined for the project: in some cases the criteria could not be met for the actual resources; in other cases future iterations of the collection may add more texts to a collection.

ELTeC extended includes fiction texts encoded in an ELTeC-compliant manner, but selected according to other criteria, such as covering a period of time, or tipping the balance in favour of representativeness, etc.

[6] https://distantreading.github.io/sampling_proposal.html

[7] On the history of Slovak literature see Richter 1979; Zajac 1996; Šmatlák et al. 2003.

> sufficiently high proportion of novels by female authors and of low canonicity. This, in turn, means the lesser-resourced languages, which again are precisely what the COST Action aims to support and which make ELTeC unique, are penalized by the composition criteria. This "diversity paradox" also results in a conflict of targets: fostering the inclusion of collections of novels in lesser-resourced languages in ELTeC as a whole, while also fostering the inclusion of marginalized categories of novels in each collection within ELTeC. This is exacerbated by a third target, namely that of maintaining the comparability of the collections. (Schöch et a. 2021)

The Slovak and Ukrainian literary scholars symptomatically joined the COST Action as its last participants in the last year of the project. And while war-torn Ukraine has managed to compile its ELTeC corpus, work on its Slovak counterpart is still in progress. A research group led by Marek Debnár of the University of Nitra has been working on a fully annotated digital full-text collection of Slovak prose up to 1951 in TEI format since autumn 2021. The interview with Marek Debnár in section two of this chapter sheds some light on the difficulties encountered by the Slovak team. These are not just of a managerial nature. Political dependence and religious strife, together with the rural character of the country – circumstances that influenced the system of literary genres and the imagery of literary works, which tended towards messianism rather than realism - make it extremely difficult to decide whether a work is a novel. The linguistic situation is likewise complicated, since the standard Slovak language was not adopted until 1843 after the Protestants lost faith in a satisfying agreement with their Czech co-religionists and reached out to the Slovak Catholics. The standard orthography was agreed on even later in the nineteenth century, while literary works considered part of Slovak culture were also written in Latin and other languages of the region. As a product of a so-called small or minor culture, the Slovak novel highlights the discrepancy between the qualities of completeness, representativeness, and balance. The scarcity of texts, due to constant repression, disturbs the relationship between the sample and the population.

ELTeC's seemingly arbitrary balance seems to favor some literary cultures – English, German or French, for example – over others, as if it represents their cultural dominance, even though it was designed with equity in mind. Let us take the example of the eligibility criteria. Since the novel is a protean phenomenon in all respects, the history of Slovak culture - centuries of political dependence and religious strife - makes the decision as to the novelistic character of a piece of narrative prose all the more problematic. The dominant tendency in Slovak fiction is the fabulous character, magical realism avant la lettre, the use of folklore, or the mythicisation of the past in the historical novel. How liberal was the concept of the novel between 1840 and 1920? Similarly, the criterion of proportionality, which ELTeC favours in relation to representativeness in order to compare different national corpora, conveys some hegemonic assumptions. It is obvious that the distribution of Slovak novels will be shifted on the time axis as is the case with the proportion of women writers. The difficult conditions under which Slovak writers worked explain the overrepresentation of clergymen, the underrepresentation of women and the virtual impossibility of meeting the requirement that only 10% of writers be represented by three works. In our case, a few writers who had secured their livelihood with the help of the Christian churches produced a relatively large number of prose texts, potential novels.

### 2.2. Interview with Marek Debnár

Marek Debnár is the principal investigator of the project "Digitálna zbierka slovenskej prózy", "Digital Collection of Slovak Prose", which is presented on the website of the Slovak Academy of Sciences as follows.

> The project is focused on applied interdisciplinary research in digital humanities. The aim of the project is to create an online accessible and comprehensively annotated full-text digital collection of Slovak prose until 1951 (with regard to copyright and GDPR). The digital collection of literary texts of this scope and characteristics is still absent in Slovakia, which is especially felt by literary scholars and the wider professional public. The collection will be processed and annotated according to European standards (ELTeC) so that it can be applied to advanced methods of computer processing of literary texts, generally referred to as "distant reading". The project also includes the education of scholars and the general public in the use of the digital collection of Slovak prose, and dissemination of scientific methods based on distant reading to the wider range of humanities. The main applied outputs of the project are following: an online digital collection of Slovak prose until 1951, an interactive web portal, user guides, and a series of educational workshops.

We asked Marek Debnár the following questions, preceded by a preamble: The political and cultural history of Slovakia puts into sharp relief the problems of eligibility, the paradoxes of representativeness, as well as the tension between representativeness and balance, or, to be precise, the fact that the apparently arbitrary balance of ELTeC is more adaptable to some literary cultures (e.g., English, German, or French) than to others, as if it represents their cultural dominance even though it was designed with equity or universality in mind. We are interested both in the technical and the research components of the project.

CLS INFRA: How advanced is the project? As I understand it, it is intended to last until 2025. Could you say something about the timeframe?

MD: The Digital Collection of Slovak Prose project began in September 2021 and will continue until December 2025. We are therefore in the first/second phase of the project, which focuses on the acquisition of texts in various formats and their conversion into xml format. This year, 2023, we would like to complete a test collection of 100 novels by December, which would also be usable for ELTeC.
General project timeframe:
Phase 1: networking and preparatory theoretical-methodological phase (July 2021 - July 2022)
Phase 2: collecting files, digitalization, and optical character recognition (August 2022 - December 2023)
Phase 3: annotating and testing the prototype of the pilot version (January 2023 - December 2023)
Phase 4: finalizing collection and prepare online access (January 2024 - December 2024)
Phase 5: full access to the collection and educational activities and dissemination of distant reading approaches (January 2025 - December 2025)

CLS INFRA: Another related question concerns the general organization of the work, i.e., the relations between the Slovak Academy of Sciences and the University of Nitra, the relations within the Academy, and between individual researchers. What are your areas of responsibility?

MD: The principal investigator of the project is me and the Faculty of Arts in Nitra and my former workplace, the Centre for Digital Humanities, where I was the director. The secondary investigator institutions are the Department of the Slovak National Corpus of the Institute of Linguistic, Slovak Academy of Sciences and the third is the Institute of Slovak Literature of the Slovak Academy of

Sciences. From the Academy we have 4 members of the team and from the faculty there were 5 members. Unfortunately, due to the last new reform of higher education and lack of funding, several departments were closed/canceled in all universities in Slovakia, and this affected my department as well. I am an associate professor and the supervisor of the literary aesthetics program at the faculty, so I was transferred as a supervisor to the Department of Journalism and New media, but my colleagues with PhDs must have been let go. This complicated the last year of the project considerably. The Center for Digital Humanities at the Faculty was dissolved in August 2022.

CLS INFRA: How do you go about the public policy aspects of your project, such as "the education of scholars and the general public in the use of the digital collection of Slovak prose" or the "dissemination of scientific methods based on distant reading to a wider range of humanities"? Do you postpone the realization of the scheme until the last phase of the project when the collection is more or less complete?

MD: Yes, this phase is scheduled for 2024. We are currently participating in conferences and workshops. The plan is to publish two manuals in 2025 and we are also planning a series of workshops.

CLS INFRA: Do you continue to cooperate closely with the Ukrainian colleagues (as the article Debnár & Jesypenko 2020 suggested was the case)? In contradistinction to you, they have their ELTeC github repository.

MD: My Ukrainian colleague is currently a refugee in Canada because of the situation in Ukraine. We are in contact more personally than professionally. The situation in Slovakia is also quite challenging, which is why I am currently (from the beginning of Russian war) working as an external expert for the Ministry of the Interior in the Hybrid Threat Unit, where I analyse social network data on Russian propaganda in Slovakia. This work takes a lot of my time now, but this is the current situation.

CLS INFRA: What is the proportion of texts that you must digitize in the framework of your project, and to what extent can you use the off-the-shelf general-purpose collections, such as the digital collections of the Slovak National Library (https://dikda.eu/) and the University Library in Bratislava (http://digitalna.kniznica.info/browse), the collection of the Slovak classics in HTML and rtf provided by the SME daily (https://zlatyfond.sme.sk/), the Slovak Wikisource.

MD: Our goal was not primarily to digitize, but to collect the available digital versions. We planned to digitize only special works or editions. The problem arose with resources, which are very difficult to obtain. Wikisource contains almost nothing. The Golden Fund was created by volunteers, and libraries do not have as many texts as we anticipated. This also fundamentally changes the nature of the collection. We work primarily with the Slovak National Corpus, which has a collection of prose mainly after 1958, and we were approaching the publishing house Kalligram, which published collection of Slovak literature library in cooperation with the Institute of Slovak literature at the Slovak Academy of Sciences. The plan was to use the project to acquire works of older literature and to combine them with the post-1958 works that are processed in the Slovak National Corpus. The aim was to have a synergistic effect, to create a comprehensive collection from the 19th century to the present day.

CLS INFRA: What devices / pipelines do you use to standardize the texts coming in different formats?

MD: We have conversion scripts from various formats to xml. Scripts were developed at the Slovak National Corpus, which is one of our partners in the project.

CLS INFRA: A series of questions concerning the process of compilation of the Slovak corpus in the comparative context, or, to rephrase it, the implications of the cultural situation of Slovak literature for the desired qualities of the corpus, such as the eligibility of texts, their representativeness, and balance: To start off: An important question would be whether the actual work on the collection has changed the way you view the general task in comparison with the vision presented in the article written with Dmytro Yesypenko, "Budovanie komplexných a reprezentatívnych"…? (Debnár & Jesypenko 2020)

MD: This study concludes by mentioning Slovakia's lag in the field of digital humanities and digital literary collections. It also mentions that there is currently a project submitted to the Agency for Science and Research for the creation of a digital collection of Slovak prose. This is the project we are currently discussing, so things have progressed, and we managed to obtain a grant, although they deducted 40 percent from the proposed budget.

CLS INFRA: Completeness: do you agree with some creators of ELTeC (Schöch at al. 2021) that it should be substantially more difficult to reach the threshold of one hundred Slovak novels than in the case of cultural superpowers such as Great Britain, France, or Spain? Can it be that the Slovak compilers' quest for completeness can negatively impact representativeness and balance? As for the criterion of eligibility: To our understanding, Hungarian and Austrian oppression—as every oppression hinders realism in favor of messianic and magical thinking—along with the rural character of Slovakia, which was devoid of metropolises and where the Western novel thrived, impacted the poetics of narrative prose. The dominant tendency of this fiction is its fabulous character, magical realism avant la lettre, leaning on folklore, or the mythicization of the past in the historical novel. How broad is the notion of the novel between 1840 and 1951? Should the corpus compiler favor linguistic, geographical, or morphological criteria?

MD: At this stage, our goal is to collect everything we can. The balanced corpus is a question that I addressed in a study that is currently in press. A representative corpus (selected corpus from the whole database) will be balanced according to specific Slovak criteria. We adhere to individual periods and literary movements as defined by literary history, while also distinguishing between first-person and third-person narration, among others, in the annotation of texts. It would be premature to specify the concept of the Slovak novel based on quantitative data.

CLS INFRA: Regarding eligibility again: What are the linguistic criteria for text selection? I am asking because the modern Slovak language was only codified in 1843 and it took some time to agree on an orthography (a factor that even gained prominence in machine-readable corpora). Moreover, Slovak literary history encompasses novels written in Latin and neighboring languages throughout the nineteenth century, and some Slovak novels are written in more than one language (for example, Karol Kuymány's *Ladislav* (1839)).

MD: When it comes to older texts, our primary focus is on selecting texts that were published after the codification, and we rely on edited older texts that have been published in the Library of Slovak Literature. Our collaboration with the Institute of Slovak Literature allows us to address these questions in an operational manner. However, there is another question regarding texts written in Hungarian. In

this regard, we are in a similar situation as Ukraine with literature written in Russian. Honestly, we haven't resolved this question yet.

CLS INFRA: Do you think that the inevitable conflict between representativeness and balancing becomes particularly grave in the case of the corpus of the Slovak novel? One might rightly say that the case of the Slovak novel lays bare hegemonic assumptions hidden behind the ostensibly neutral, even intentionally progressive demands of ELTeC, which for the sake of comparability of different national corpora, gives preference to proportion in relation to representativeness. However, the criteria the compilers strive to meet shed some light on the irreflexive self-reliance of the established *Kulturnationen*.

MD: Yes, this conflict is significant. The criteria used for major European literatures such as French or German cannot be directly applied to smaller literatures like Slovak. In fact, we don't even have the required number of novels for the ELTeC preparation. Nevertheless, we will strive to deliver the Slovak part of ELTeC by the end of the year. As for our own collection, we have to adapt the criteria; otherwise, the collection would not be usable.

CLS INFRA: It is obvious that the distribution of Slovak novels will shift on the time axis.
Why is the terminus *ante quem* of the project 1951? Does the selection of this end point in time mean that you no longer wish to become a part of ELTeC? How, if at all, do you deal with the proportional distribution of texts between the periods 1840 – 1859, 1860 – 1879, 1880 – 1899, 1900 – 1920, characteristic of ELTeC? What is the starting date for the texts (*terminus post quem*)? What is the oldest text selected for the collection?

MD: The starting date is 1840 because it is related to the codification. Our intention is to be part of ELTeC with works that meet the criteria. The year 1951 is related to the copyright law in force in Slovakia and is only approximate. In fact, in the latest version, it has been shifted as copyright lasts for 70 years. Our goal is to connect it with the corpus of contemporary fiction from 1958 (the introduction of the currently valid language norm) to the present, which is being developed at the Slovak National Corpus department. By combining them, we would be able to create a relatively comprehensive collection from 1840 to the present.

CLS INFRA: What about the proportion of women writers?

MD: The representation of female authors is very low, especially in the 19th century, but it has been consistently increasing since the 20th century.

CLS INFRA: What about the condition that only 10% of writers be represented by three works? Do you believe that proportionally fewer people were able to write and PUBLISH novels than in, say, Germany?

MD: This is a question that concerns the literary canon in the case of ELTeC as well as domestic literary production. It is evident that publishing opportunities in Slovakia were more limited compared to Germany. Many works were published in magazines and did not take the form of printed books, or the book editions had a second printing only if the work was successful.

CLS INFRA: How do you perceive the relationship between your collection and the literary canon? I mean both the established canon of Slovak literature and the place of Slovak literature in

European consciousness. Do you believe that distant reading and multilingual tools and pipelines can somehow help include the Slovak novel in the history of European literature?

MD: We are firmly convinced that the inclusion of the Slovak novel in European literature in a digital form, using distant reading tools, will have great significance. Our literary research has long been focused solely on our own provenance, which is evident even in the realm of close reading. To see Slovak literature in a broader context is essential for preserving its reflection.

Regarding the broader context: The utilization of distant reading tools will enable us to explore the Slovak novel in relation to European literature, uncovering new perspectives and fostering a deeper understanding of its value and contributions. This inclusion will not only enrich the European literary landscape but also provide insights into the unique aspects of Slovak literature and its cultural significance.

CLS INFRA: Our wish for Slovak literature on the international stage is certainly the same. Thank you for the informative interview.

### 2.3. Problems of Corpus Compilation

In the case of the Slovak novel, the compiler of the corpus can refer to a number of high-quality general-purpose collections, such as the digital collections of the Slovak National Library (https://dikda.eu/) and the University Library in Bratislava (http://digitalna.kniznica.info/browse), both affiliated with Europeana, and the collection of the Slovak classics in HTML and rtf provided by the SME daily (https://zlatyfond.sme.sk/autori#U). Of course, there is also the Slovak Wikisource.

An SPARQL query in Wikidata looking for Slovak authors[8] renders only 64 entries (as of 02.03.2023):

```
SELECT ?PersonA ?PersonALabel WHERE {

?PersonA wdt:P172 wd:Q171336.
?PersonA wdt:P106 wd:Q36180.

   SERVICE wikibase:label { bd:serviceParam wikibase:language "en". }
}
```

In comparison, there are 140 Polish authors and as many as 739 Czech ones accounted for in the knowledge graph.

[8] The caution is needed when drawing conclusion sfrom this query. The porperty wdt:P172 “ethnic group” is marked as "disputed"; therefore problems may occure in connection with the reluctance of many contributors to to attach this property. If we use ustead Property wdt: 1412 “language spoken, written or signed” we arrive at 1235 whose occupation (P106) was writer and who used the Slovak language.

Even more strikingly, the SPARQL querry for the Slovak Novel provides only four results.

```
SELECT * WHERE {
  ?novel wdt:P7937 wd:Q8261 ;
        wdt:P407 wd:Q9058 .
}
```

In the search for potential material for a corpus of the Slovak novel, we have decided to rely on the provisional catalogue of corpora curated in the CLS INFRA project, in particular by Work Package 6.[9] This catalogue will be the basis for the Corpus Exploration Platform tool (see section 3.3 and 4.4). Two research assistants, Elisabeth Bauer and Annika Blietz, manually searched for Slovak novels in the text collections listed in the catalogue. They started with the search for texts in Slovak language.

| **multi/mono language corpus** | | Slovak language |
|---|---|---|
| ÖNB Digital / Österreichische Nationalbibliothek | https://onb.digital/search/543714 | 820 Slovak records (3364 Russian, 29475 English, 2062162 German records) |
| Europeana | https://www.europeana.eu/de/search?page=1&qf=LANGUAGE%3A%22sk%22&query=&view=grid | European heritage; including Slovak (11069 records) |
| Polona | https://polona.pl https://polona.pl/search/?filters=language:słowacki,public:1,hasTextContent:0 | Slovak: 157/2050736 records online, 1230/3731749 records in total |
| DIKDA – Písomné kultúrne dedičstvo Slovenska | https://dikda.snk.sk/search?languages=slo | 840340 Slovak entries,938110 total (DIKDA, the digital library of the Slovak National Library (SNL)) |
| Zlatyfond SME 2.0 | https://zlatyfond.sme.sk/dokument/o-projekte/ | classic Slovak literature, digitalised |
| Univerzitná Knižnicsa v Bratislave | http://digitalna.kniznica.info/browse | University library of Slovak literature, monographs, maps, hand writings etc. (listing 104723 records) |
| Slovak literature corpus | https://bonito.korpus.sk/run_guest.cgi/first_form | 1155742085 records |

(Table 1: Corpora containing Slovak texts)

[9] https://zenodo.org/record/7520287#.ZGfDQnZBy70

A general examination of the first 86 entries in the corpus catalogue shows that works of Slovak literature are underrepresented. Even the corpora that cover the most languages (e.g. Oxford Text Archive, Gutenberg Project) do not list any Slovak literature, although related Slavic languages such as Slovenian or Serbian are available. For example, as far as Slovak texts are concerned, Europeana only contains newspapers from the First World War and one manuscript from the same period.[10] Where Slovak texts do appear, the number of records is significantly low - a situation that is reflected in the third column of the Table 1. Of course, Slovak literary texts are available in monolingual corpus of Slovak origin.

The next step of the data landscape consisted in a detailed examination of the Slovak language corpora, which would determine which retrievable texts would be suitable as novels. Below we list our preliminary results.

DIKDA – Písomné kultúrne dedičstvo Slovenska

DIKDA as a text collection curated by the Slovak National Library has an excellent filter.
Below you will find the results for 'Slovak literature' and 'Slovak novel'.
https://dikda.snk.sk/search?keywords=Literat%C3%BAra%20slovensk%C3%A1,,Slovak%20literature,,Slovensk%C3%A1%20literat%C3%BAra&licences=only_in_library&genres=Slovensk%C3%A9%20rom%C3%A1ny&doctypes=monograph&sort=alphabetical

1) Ako divé husi - román o pamäti, (Ferko, Vladimír)
https://dikda.snk.sk/view/uuid:83a90d55-8bf6-4709-8d2c-fa7c7f034c61?page=uuid:07a40428-5340-464f-bca6-78155b144b56
2) Anna Šarišská amazonka (in two versions), (Šranková, Eva Ava)
https://dikda.snk.sk/view/uuid:82417da7-272b-4004-9bc2-7964e361ac8e?page=uuid:7a5f1492-a1ae-4d95-a97f-be957349a2b1
3) Buď vôľa tvoja 4. Voľný pád, (Belangelo, Sun)

[10] 1) newspaper (WWI): Slovenské noviny pre politiku, spoločenský život, priemysel a hospodárstvo
2) manuscript: Krvavé sonety
3) newspaper (WWI): Národné noviny
4) newspaper (WWI): Slovenské ľudové noviny orgán slovenskej ľudovej strany
5) newspaper (WWI): Národní hlásnik noviny pre slovenský ľud
6) newspaper (WWI): Dennica ženský list pre poučenie a zábavu
7) newspaper (WWI): Slovenský týždenník
8) newspaper (WWI): Domácnosť a škola časopis rodinný a učiteľský
9) newspaper (WWI): Domácnosť a škola časopis rodinný a učiteľský
10) _,,_ : Robotnícke noviny slovenský orgán Sociálnodemokratickej strany Uhorska
11) _,,_ : Evanjelický posol zpod Tatier
12) _,,_ :Slovenský denník
13)_,,_: Vlasť a svet poučno-zábavný a hospodársky časopis
14) _,,_: Veselé noviny humoristický mesačník
15) _,,_: Rarášek humoristicko-satyrický mesačník k obveseleniu a zábave
16) _,,_: Posol božského srdca Pána Ježiša časopis braterstva najsvätejšieho Srdca Ježišovho
17) _,,_: Sborník Museálnej slovenskej spoločnosti
18)_,,_: Právny obzor teoretický časopis pre otázky štátu a práva
19)_,,_: Duchovný pastier revue pre teológiu a duchovný život

https://dikda.snk.sk/view/uuid:c30b5dce-76b5-4abb-b258-fe57f656c286?page=uuid:b3847a6e-f0d7-487c-8db9-ed82ee0098d6
4) A cincea corabie, (Kompaníková, Monika)
https://dikda.snk.sk/view/uuid:9e92c591-f3bc-49c4-bdb2-da9044d39629?page=uuid:080b2a32-0908-435d-a50c-5af29e45e2e6
5) Čachtická paní (Kompaníková, Monika)
https://dikda.snk.sk/view/uuid:9ec03cf8-36c7-42ee-b1a4-0f42cd6c3720?page=uuid:ec8e5f5f-1dbc-406d-be61-3069c81a0335
6) Čas nezastavíš (Gregorová, Hana)
https://dikda.snk.sk/view/uuid:78630dbc-495e-4e67-93ca-77566efc5596?page=uuid:aa82c6e9-8ef6-4c1d-84b6-8298f922f2e6
7) Dielo (Šikula, Vincent)
https://dikda.snk.sk/view/uuid:842b85f5-2385-4a9a-a82a-7112efcff982?page=uuid:953a1fb5-6bdf-4805-bd80-72b42bd86a3a
8) Doktor Jessenius (Zúbek, Ľudo)
https://dikda.snk.sk/view/uuid:3cc2bc53-2e06-446d-880a-d0b73f72ba1f?page=uuid:050150a5-de29-4500-812e-7fbb9a510298
9) Dva (k)roky ke štěstí (Holiga, Jaroslav)
https://dikda.snk.sk/view/uuid:0fd33936-436f-4f64-bfdb-d06b107f4310?page=uuid:99d3237b-494a-4574-90ab-0e670faafeca
10) Hlas krvi (Fábryová, Blanka)
https://dikda.snk.sk/view/uuid:ada306f1-0eed-4435-a503-854a3c01567c?page=uuid:f83b237b-5388-4d41-8e61-5e9d503000c7
11) Hrdinovia (Hutár, Jozef)
https://dikda.snk.sk/view/uuid:218ebd39-f92b-41b1-ab20-c41af2833c62?page=uuid:1d040abb-a3cd-4d51-b505-ca3ba8368e83
12) Hriech (Rákay, Anton)
https://dikda.snk.sk/view/uuid:513e2dac-e42c-4b66-801a-3a5c87367e4a?page=uuid:4ce896fb-733f-449b-9b12-b1c7916c05c2
13) Jánošík (Urban, Milo)
https://dikda.snk.sk/view/uuid:dd739bd7-8ff2-4f9d-933c-3b2c7fba7ff9?page=uuid:0c8f8ecf-5985-4d7b-a4da-75d1ebd221ba
14) Jej Americká svokra (Sinak, Maria K.)
https://dikda.snk.sk/view/uuid:b39e7a01-0b0a-4acf-a76e-b438629ec64a?page=uuid:c1830238-2730-4aa2-a295-4e17237437b6
15) Krv (Sloboda, Rudolf)
https://dikda.snk.sk/view/uuid:b735edf9-59e2-4473-b676-38f778a7e273?page=uuid:d35793d6-26c0-4108-a914-47a2f19519f4
16) Luč (Kužel, Dušan)
https://dikda.snk.sk/view/uuid:92642731-c32d-4103-b77a-08388fb9cbab?page=uuid:c48065ec-7362-4de6-9429-954ee296fb3f
17) Majki (Rankov, Pavol)

https://dikda.snk.sk/view/uuid:6ee03abf-2144-4372-8b8b-0f756e7684d1?page=uuid:af4eac26-4526-4e66-91b0-431c65c4d087
18) Marmota (Platko, Igor)
https://dikda.snk.sk/view/uuid:a3cfed5d-72fb-405d-8a22-fa21ff38182c?page=uuid:7a2e66b9-85b4-47b9-bad7-8144e599d40e
19) Maroško (Rázus, Martin)
https://dikda.snk.sk/view/uuid:cefa97f9-aaf4-4cd5-9541-4c3952668bee?page=uuid:e8225ba8-cffc-436e-bbfe-c21c16060b87
20) Môj skvelý brat Robinzon (Blažková, Jaroslava)
https://dikda.snk.sk/view/uuid:2f1233a8-90ea-4ab0-b588-961b42f2001b?page=uuid:4db4bf68-4132-4afd-b839-c98918275674
21) Nežný dotek nenávisti (Hrašková, Eva)
https://dikda.snk.sk/view/uuid:b6368e51-200d-4aef-bee1-1f5c51e151c0?page=uuid:bde50d5a-90d0-41cb-bc94-576ceae8e2b4
22) Nôž (Thal, Juraj)
https://dikda.snk.sk/view/uuid:d071337d-d3ec-473a-9133-c54b969bba83?page=uuid:93c4033f-edde-414b-8d4e-5d7e640f3a05
23) Piata loď (Kompaníková, Monika)
https://dikda.snk.sk/view/uuid:6ffa31eb-987f-4762-8965-2b3f085bc99a?page=uuid:87bbb6cd-fd07-4420-be73-982a4b50e1c3
24) Pole neorané (Jilemnický, Peter)
https://dikda.snk.sk/view/uuid:dd316d63-a6ce-4e47-a46c-d56ee6ff5f0d?page=uuid:bf58f400-21ed-4514-b696-f00c35b9ffca
25) Povedal, že budem šťastná (Gregorová, Adriana)
https://dikda.snk.sk/view/uuid:9e76611c-a345-42af-a031-d8a337309b53?page=uuid:764d1c50-59ea-4145-b13f-00ee0c3ca8f5
26) Prázdniny na tretiu (Majchráková, Svetlana)
https://dikda.snk.sk/view/uuid:1afd8ef4-9b4e-4491-be0e-59bb872a22b6?page=uuid:f1c9702f-8d69-4798-9196-18ca73ed8939
27) Priznání (Hlavatá, Dana)
https://dikda.snk.sk/view/uuid:21939156-c12a-411b-8862-526cd19d6444?page=uuid:c778e752-949b-4262-8f45-2503f0763e17
28) Rockové tango (Wurm, Monika)
https://dikda.snk.sk/view/uuid:68f46c0d-b0ef-41fc-b882-b90ca5435bc9?page=uuid:d65eedf5-58ce-489d-a43b-18191e1d46f9
29) Sladké časy (Lachkovič, Adolf)
https://dikda.snk.sk/view/uuid:5c06b13e-d58c-4579-8aad-51e252c11afd?page=uuid:8be3aa44-01d7-4b0a-977b-342cf5581caa
30) Sluči se na părvi septemvri (ili v brug den), (Rankov, Pavol)
https://dikda.snk.sk/view/uuid:7c1cac86-23e2-4dc4-9ca3-8f066f9b68db?page=uuid:044d37fd-c0f8-4771-99ec-372d3dbebf28
31) Slzy prodaných dívek (Wurm, Monika)

https://dikda.snk.sk/view/uuid:80ac031e-e200-4e5c-aeac-56bf3d66f3e2?page=uuid:79c48938-f2d8-4739-9dac-a7057df11ff4
32) Strach (Karika, Jozef)
https://dikda.snk.sk/view/uuid:a93fd50b-31cc-4333-a6f1-edb82f0cfc4e?page=uuid:ac83bb24-0c07-49d2-bceb-16bc466a15a2
33) Svetlo pod halenou (Klas, Peter)
https://dikda.snk.sk/view/uuid:7289c914-b37d-48e7-a547-67da161fe3d6?page=uuid:3d79ba7b-a729-48fd-bc75-509d4b449c07
34) Šťastie za dverami (Ferko, Miloš)
https://dikda.snk.sk/view/uuid:b2046d96-7e83-48f4-ab7f-6fee54b72d18?page=uuid:6e486390-ef25-4d01-aa85-7e27a5c41a6e
35) Tantalópolis (Macsovszky, Peter)
https://dikda.snk.sk/view/uuid:fa110c8c-ecc1-48eb-ada7-c330516b0eff?page=uuid:b714e6ec-ebb8-46db-b042-831e2da5ba8c
36) Tieň radosti (Holiga, Jaroslav)
https://dikda.snk.sk/view/uuid:5626ecaa-ebb1-4ee6-a148-42c5ceb9a667?page=uuid:81148a1d-425e-4275-a6a2-5c7a10dce57a
37) Trpaslíci (Lazarová, Katarína)
https://dikda.snk.sk/view/uuid:2bd99ec2-f5b8-4f58-8ba5-11dadd091574?page=uuid:f8ecba29-2c9c-41fb-8071-dfd35bddfe8f
38) Úsmevy a chmáry (Dováľ, Štefan)
https://dikda.snk.sk/view/uuid:e8383356-ba9e-4034-8725-bd080b52e560?page=uuid:7cdb4a5c-fe45-42d9-88fa-f21845676973
39) Zákonodarci (Šándor, Elo)
https://dikda.snk.sk/view/uuid:a60631ac-39d3-4935-9df7-bf66b035ce9d?page=uuid:1c177da5-ee4a-41fc-981f-d5600641c795
40) Zánik zeme Magog ; Stena (Dančo, Lukáš D.)
https://dikda.snk.sk/view/uuid:30b32bc5-89cd-4c01-8389-4505857abb05?page=uuid:18f1f479-e3fb-4cba-9619-f9410140d8cb
41) Zobrané spisy (Hronský, Jozef Cíger)
https://dikda.snk.sk/view/uuid:869c1a3e-a6dc-4dc4-a211-e3c270740e46?page=uuid:681ddc57-22e3-4a36-8e1b-4bbf68ecaabd
42) Zrada (Macháčová, Adriana)
https://dikda.snk.sk/view/uuid:4397524a-c316-4a71-b192-d756eeefa6aa?page=uuid:1057a654-91a9-4f5c-a3b1-fc8dd9145bca
43) Železom po železe (Urban, Milo)
https://dikda.snk.sk/view/uuid:f72d6b8c-a95f-421d-ad5b-5837c2d6d346?page=uuid:f7ec34ca-29c4-44af-8991-9f40570b62d9

Zlatyfond SME 2.0 (the Golden Collection of the SME Daily)

Link: https://zlatyfond.sme.sk/autori
This collection, the richest in novelistic material, has almost no metadata. It is often necessary to consult external sources (handbooks, Wikipedia, etc.) to find out whether a prose text is classified as a novel. The corpus is organised by author. For each author there is an entry, which sometimes helps to find out whether an author wrote novels. Below are the preliminary results with all the candidates for novels and thus for our ELTeC corpus. We found the following candidates for novels, each of which should be now tested either in the process of reading or by referring the title to a database, which can contain information on the genre. Unfortunately, Wikidata, as we have seen, does not provide much information on Slovak literature so we are left to our own devices and forced to rely on our experience as readers.

Jozef Ignac Bajza: Príhody a skúsenosti mládenca Reného.

Koloman Banšell: Atalanta, Šuhaj ako z iskry, Na parolodi a železnici, Okyptenec, Túhy mladosti, Z tých krajších časov

Ján Bežo: Perly pre našu remeselnícku mládež

Anton Bielek: Keď sa starí začnú hašteriť, Na mylných cestách, Na salaši, Obecné počty, Obrázky z hôr, Poviestka z hôr, Pre cudzie viny, Rozprávka z Poľany, Z dôb utrpenia

Juraj Bindzar: Hekuba

Samuel Bodický: Žena dvoch mužov, Nihilizmus, Nová kométa, Opatrná matka, Otcova kliatba, Prvá ľúbosť, Vo väzení, Zaľúbenec, Zlý duch, Zmŕtvych

Samo Cambel: Moje otcovstvo, Pavlus, Pochybenie, Prekazená služba, Slečna Nina, Stará matka, Trest, V snahe, Vytriezvenie, Z tmavej zápače, Záhorcovo šťastie, Zlomená duša

Jolana Cirbusová: Žobráčik, Bez roboty, Cez zatvorenú hranicu, Kožušníci, Len odišiel a nevrátil sa, Najkrajšia, najmúdrejšia, najbohatšia, Pred večierkami, Prispôsobenie, Tulákove radosti, V tieni smrti

Ján Čajak: Únos. Obrázok z dolnozemského života, Špitál, Bako a Miko, Cholera, Chudobný boháč, Ecce homo!, Fuksi, Harun Arrašid a indický cár, Ja si môj mliečnik nenačnem, Jeden deň zo, života vlasteneckého kňaza, Jedna korunka, Jožkova svadba, Len pekne, Malí zbehovia, O jednom otcovi a dvanástich synoch, Pokuta za hriech, Pred oltárom, Predaj hory, Pytačky, Rodina Rovesných, Rozpomienka na veľkonočnú oblievačku, Strýc Miško, Tridsaťpäť dekagramov, Ujčekov Mikušov posledný deň, V druhej triede, V tretej triede. Črty, Vianočné dojmy, Vohľady, Vtáčie hniezdo, Z povinnosti, Zo života malého nezbedníka, Zuza

Ján Červeň: Divé vtáča, Dva slnečné dni, Modrá katedrála, Prorok, Sedem listov, Svätá žiara, Z denníka, Zlomený kruh

Jonatán Dobroslav Čipka: Poklad

Ferdinand Dúbravský: Bohyne na Žítkovej, Ján Velflin z Nepomuk, Krvavý hostinec a iné povesti, Malé Karpaty a Biela Hora, Nepodarilo sa, Peštiansky ideal, Pijavice, Podozrivá nevesta, Skalica, Slovenská kuchárka mlsnej tetky Feriny z Dúbravky, Uhorská Skalica

Olga Feldekova: Poviedky

Mikuláš Štefan Ferienčík: Brüder, Irma, Lehrer in Jedlovsk, Unruhige Ruhe

Ľudovít Gaspar-Zaosek: Škaredá streda po fašiangoch, Čudná pani, Čudný pán, Cestovanie chorého Zaoska do bardiovských kúpeľov v Šáriši, Chystanie sa jednej familie na horniacky kaliko-bál, Dva listy, Dva listy. Juraj Rebro na vojne, Hádka Zaoska s Cvajfúsom, Konferencia pri káve u starej, Zaoskovej, Môj životopis, Mrzutosti na ceste, O jazyku človeka, Panský oddiel z panorámy pekla, Ponosy učiteľa Gilisztaya, Rozpomienka na jeden trnavský piknik, Tri kapitole o biedach krčmára, Trnavský bál, Viedeňské zkúsenosti sedliaka Zaoska, Vykladanie karat, Zaosková cesta ako novinkára na pokojovú konferenciu do Haagy

Alžbeta Göllnerová-Gwerková: Žena novej doby, Zsigmond Móricz: Zatratené zlato I, Zsigmond Móricz: Zatratené zlato II

Josef Hofer: Kopaničářské povídky

Martin Hranko: Burko, Furko a Murko, Ježkovci, Kukučka, Vrabčiak Ťulko

Jozef Miloslav Hurban: Gottschalk, Gottschalk, Korytnické pohárik alebo Veselé chvíle nezdravého človeka, Od Silvestra do Troch kráľov, Olejkár

Ján Chalupka: Bendeguz, Gyula Kolompos a Pišta Kurtaforint. Donquijotiáda podľa najnovšej módy.

Václav Chlumecký Enšpenger: Drobné humoresky, grotesky, alegórie a besednice, Janko Kosák v nebi, Profesor Šiltao

Dobroslav Chrobák: Kamarát Jašek, Drak sa vracia, Fejtóny

Elena Ivanková: Štedrý večer, Žlto-brokátový plášť, Červená blúza, „Počiatok románu“, Barko Mári, Boj so šarkanom, Citróny, Dcéra pána Karabína, Dobrá partia, Dvom pánom slúžiť, Glosy, Keby nebola vojna…, Kleopatrin peniaz, Koniec, Lúčenie, Menuet, Moderná bájka, Na mena,

Posledná idyla. Obrázok z doby rokoko, Pred sto rokmi, Prostá historia, Svadobná noc, Svedok lásky, Taký je život…, Trio…, Tuberóza, V krinolíne, Všechsvätých na vojne

Janko Jesenský: Zo starých časov

Peter Jilemnický: Kus cukru, O dvoch bratoch, Pole neorané, Víťazný pád, Zuniaci krok

Ján Kalinčiak: Bozkovci, Bratova ruka, Knieža liptovské, Láska a pomsta, Milkov hrob, Mládenec slovenský, Mních, Orava, Púť lásky, Poslední počestnost spící dýmky, Rozpomienky na Ondreja Sládkoviča, Serbianka, Svätý Duch

Andrej Kalník: Poklad Inkov

Ludovít Kubani: Čierne a biele šaty, Emigranti, Hlad a láska, Mendík, Pseudo-Zamojski, Suplikant, Valgatha

Martin Kukučin: Ako sa kopú poklady, Úvod k vakáciám, Štedrý deň, Čas tratí — čas platí, „Nie milí, nie drahí“, Bacúchovie dvor, Baldo & Comp., Bohumil Valizlosť Zábor I, Bohumil Valizlosť Zábor II, Bohumil Valizlosť Zábor III, Cestopisné črty. Rukopisy, Dedina v noci, Dedinský jarmok, Dedinský roman, Deti, Dies irae, Do školy, Dojmy z Francúzska. Črty z ciest, Dom v stráni, Dosiaľ nepublikované práce, Drotárove Vianoce, Dve cesty, Hajtman, Hody, Keď báčik z Chochoľova umrie, Keď bol Matej malý, Klbká, Košútky, Komasácia, Koncepty nepublikovaných rukopisov, Koniec a začiatok, Listy priateľom a známym, Lukáš Blahosej Krasoň, Mať volá I, Mať volá II, Mať volá III, Mať volá IV, Mať volá V, Máje, Mišo, Mišo II., Mladé letá, Na ľade, Na hradskej ceste, Na jarmok!, Na obecnom salaši, Na Ondreja, Na podkonickom bále, Na prielohu, Na stanici, Na svitaní, Neprebudený, O Michale, Obecné trampoty, Obeta, Od trištvrte na osem do ôsmej, Panský hájnik, Parník, Pán majster Obšíval, Po šiestej hodine, Po deviatich rokoch, Pod vládou ženy, Poza školu, Pozor na čižmy, Praha v znamení výstavy, Prečo Adam Chvojka spáva teraz už doma, Prechádzka po Patagónii I, Prechádzka po Patagónii II, Prechádzka po Patagónii III, Prechádzka po výstavisku, Prechádzka po Vlkolíne, Prechádzky po výstavke v Prahe, Pred pekný domec. (Neúplný koncept), Pred skúškou, Preháňanky, Prvá zvada, Punta Arenas, Regrúti, Rijeka — Rohić — Záhreb, Rodina, Rohy, Rozmajrínový mládnik, Rysavá jalovica, Slepá kura a zrno, Spod školského prachu, Susedia, Svadba, Sviatočné dumy, Syn výtečníka 1, Syn výtečníka 2, Teľa, Tichá voda, Tiene i svetlo, Tri roje cez deň, V Dalmácii a na Čiernej Hore, Výlety, Veľkou lyžicou, Vianočné oblátky, Visitatio canonical, Z lovcových zápiskov, Z našej hradskej, Z teplého hniezda, Za ženou, Zakáša — darmo je!, Zádruha, Zápisky zo smutného domu, Zo študentských časov, Zo stupňa na stupeň

Kurd Lasswitz: Ľudia z Marsu

Kálmán Mikszáth: Gavalieri, Hluchý kováč Prakovský

Milan Thomka Mitrovský: Cyntia, Intonácie jari, Lipka, M. Th. M, Maiestas artis., Umenie, veda a literatúra, Pani Helène, Pelikán, Sopra Minerva, Z listov bratislavským felibrom, Zo zápisníkov

Ladislav Nádaši-Jégé: Adam Šangala, Aké sú naše cesty?, Ako sa dievčatá vydávajú, Ako sa mladí ľudia ženia, Alina Orságová, Annalena, Žart, Česť, Články, Babylonská veža, Biela labuť, Boj o pokrok, Brok, *Bubuš, Bubulík a Bubulinko*, Cesta životom, Cholera, Daromná robota, Dáždnik, Dáma v hoteli Bellevue, Deľba, Dedinský notár, Dlhá chvíľa, Dobrý deň, Dr. Garvič, Druhou triedou, Duch môjho strýka, Dve vdovy, Eh, Katka!, Gemma docci, Honba, Horymír, Hriešne hnevy, Hriešni ľudia, Jaríkovský kostol, Jedovaté kvety, Kačenka, Kapitalista, Katka volí, Kúra, Kozinský mlyn, Krásne časy rokoka

Anton Ottmayer: Úprimná láska, Krajina šťastia, Stála láska, Tajomná láska

Viliam Pauliny-Tóth: Trenčiansky Matúš

Karel Poláček: Dům na předměstí

Gustáv Reuss: Grófka Mária Betlenová, Hviezdoveda

Kristina Royova: Abigajil Karmelská, Ako Kvapôčka putovala, Ako prišli lastovičky domov, Ako zbohatnúť, Ako zomieral Sláviček, Šťastie, Šťastlivé Vianoce, Šťastní ľudia, Šťastný Štedrý deň, Bez Boha na svete, Deti hauzírerov, Divné milosrdenstvo, Druhá žena, Kde bol jeho otec?, Keď nikde nebolo pomoci, Keď sa život začínal, Kráľovná zo Sáby, Lótova žena, Moc svetla, Na rozhraní, Navrátený raj, Náman Sýrsky, Opilcovo dieťa, Otcovrah, Peterko, Prišiel domov, Slnečné dieťa, Sluha, So svetlom, Splnená túžba, Staniša, Stratení, Susedia, To, čo večne trvá, Traja kamaráti, Tuláci, V pevnej ruke, V slnečnej krajine, Výstražný sen, Vo vyhnanstve, Za presvedčenie, Za svetlom, Za vysokú cenu

Terezia Vansova: Šapšanko, Čo komu súdené, Biela ruža, Boženka, Chovanica, Danko a Janko, Divočka, Hojže Bože!, Ján Vansa (Výber), Johankin zajac, Julinkin prvý bál, Kar, Kliatba, Magdalena, Matky, Milka — mašamódkou, Milku dajú na edukáciu, Nové šatočky, Obete márnomyseľnosti, Ohlášky, Paľko Šuška, Pani Georgiadesová na cestách, Pani veľkomožná, Púť za šťastím, Prsteň, Recepty prastarej matere. Nová kuchárska kniha, Redaktorské skúsenosti, Rozsobášení, Sestry, Sirota Podhradských, Stará pieseň, Supplikant, *Terézia Medvecká, rodená Langeová,* Vlčia tma, Z našej dediny

Jonáš Záborský: Šofránkovci, Žihadlice, Borzajovci, Chrňo und Mandragora, Frndolína, Jurát, Kulifaj, Mroč, Mrzutá, Nálezca pokladu

To sum up, this corpus contains by far the largest number of Slovak novels. However, there were some difficulties when searching for Slovak novels in the corpus. First of all, the Zlatyfond SME 2.0 corpus is sorted alphabetically by author. Many genres are included, especially short prose

(short stories), poetry, drama, non-fiction, and novels. However, the genre is nowhere directly annotated and it is not possible to filter the corpus through a search. This makes it difficult to find specific genres, as you have to click on each entry. Therefore, this search is not as precise as it could have been. In order to structure the search, the research assistant relied on the biographical information of a given author. It would be beneficial to refer to a specific database that has information on the genre. Unfortunately, as we have seen, Wikidata does not offer much information on Slovak literature, leaving us to rely on our own reading experience. The search took a while, and future users would benefit from a filter option or more detailed annotation of the entries.

Univerzitná Knižnicsa v Bratislave (The University Library of Bratislava)

Link: http://digitalna.kniznica.info/browse

Filter: literature, novel

Results: 3

Plus the “Slovak” filter

Results: 1

http://digitalna.kniznica.info/zoom/85874/view?page=1&p=separate&tool=info&view=0,5,1565,2627

Title:
Anti-Fándly aneb Dúverné Zmlúwánj meziTheodulusem, tretího Franciskánúw rádu bosákem, a Gurem Fándly, Naháckim Farárem, o,a proti geho Dúwernég Zmlúwe mezi Mníchem, a ďáblem

Author: Bajza, Jozef Ignac

This corpus has apparently good and precise search filters, but very few entries of Slovak novels, which means that either the collection is unbalanced or unrepresentative of Slovak literature, or there is a problem with linking metadata to textual data. The corpus seems to be biased towards political writings as well as journalistic artefacts such as magazines and newspapers.

ÖNB Digital / Österreichische Nationalbibliothek

Link: https://onb.digital/search/543714

ÖNB Digital contains 820 Slovak records, 505 of which are newspapers. We were able to find some prosaic works. The word count would help us determine whether the are suitable for the ELTeC corpus.

1) Slovenskje povesti. (Slovakische Erzählungen.), (Rimauski, Johann)

https://onb.digital/result/1086EBEF

2) Pametnosti Bekes-Cabanske, (Haan, Lajos)

https://onb.digital/result/10608B45

3) Tatranka. Spis pokracugjcy rozlicneho obsahu, pro vcene, prevcene y nevcene ; Pracy a nakladem Girjho Palkowice (Palkovic, Georg)

https://onb.digital/result/10B0A6AB

4) Celorocne nedelne kazne ... (Hulyak, Jozef)

https://onb.digital/result/10AA244B

5) Cesta Slowaka ku Bratrum Slowanskym na Morawe a w Cechach, (Hurban, Jozef Miloslav)

https://onb.digital/result/10A5148A

6) Slovenski Stajer. Dezela in ljudstvo (no author given)

https://onb.digital/result/108B02AA

7) Slovenske povesti. Vydavaju August Horislav Skultety a Pavel Dobsinsky (Skultety, August Horislav and Dobsinsky, Pavel)

https://onb.digital/result/1049EC75

8) Pokuta za Hrjech po celych Uhrjech (no author given)

https://onb.digital/result/1096975F

9) Palecek ... Sebranj, od Jana Hybla (Hybl, Jan)

https://onb.digital/result/1072A399

10) Iwan Dusarow, Janko Kalina, nesstastny opilec. (Mally, Jan)

https://onb.digital/result/104B5C05

After sorting the different texts and books in the digital corpus of the ÖNB, it can be said that the Slovak novel is represented below average. Most records concern newspapers. The second most frequent category of texts concerns religion, such as edifying works, prayer books and religious texts and sermons. Historical and political works appear almost as often as religious ones. We had difficulties deciding whether to classify the epic as a kind of Slovak novel, and whether collections of stories and songs could be counted as novels. In the end we decided against it.

Polona
https://polona.pl/

The Polish library Polona contains 49 Slovak book records. The following are potential candidates for a novel:

1) O literatúre a o l'uďoch: (rozpomienky a úvahy), (Kalinčiak, Ján)
2) Milkov hrob (Kalinčiak, Ján)
3) Pani Helène (Mitrovský, Milan Thomka)
4) Neprebudený (Kukučín, Martin)

5) Drak sa vriacia : rozprávka (Chrobák, Dobroslav)
6) Dom v stráni (Kukučín, Martin)

The corpus does contain some Slovak novels, or so it seems. However, most entries must be categorised as non-fiction. Documents about historical events and biographies prevail. Again, the problem with classifying fairy tales and legeds arises, especially taking into consideration the general fantasy character of the Slovak prose that coincides with literature's function in a subaltern society. So, the total number of Slovak novels could be expanded or reduced depending on the stringency of the the categories.

### 2.4. Conclusion and Proposed Solutions

We conclude this section with a testimony from research assistant Annika Blietz, followed by our joint recommendations.

> After going through all the corpora including Slovak entries I can conclude that the Slovak novel is represented below average both in multilanguage and specifically Slovak corpora. However, some interesting conclusions can be drawn from the research.
>
> The first corpus I examined was the ÖNB (Austrian National Library). Here, most of the texts were newspapers. The most common topic regarding books was religion, be it sermons, prayer books, or works on religion. This was followed by non-fiction books and essays on history and politics. I also decided not to include epic poems, fables and fairy tales or collections of stories, and songs in the same category as Slovak novels. The research was not easy and the result was probably not so precise. This was due to the lack of filter options for the research and the lack of annotations to the entries.
>
> The second collection I looked at was Europeana. Here the search was quite fast due to the exact and many filter options. However, the results for "Slovak" and "text" were only newspapers from the time of World War I. In this corpus there are good filter options, but the variety of data needs to be expanded.
>
> By far the best corpus in terms of search options was the DIKDA corpus. Precise filter options lead to clear results. However, the diversity of these results was not the same. The novels were mostly from the same time period, i.e. recent literature from the last 50 years.
>
> The corpus with the most entries of Slovak novels was the Zlatyfond SME 2.0. There are many novels from different periods, sorted alphabetically by author. However, it took by far the most time to browse. The corpus has no filter options or clear annotations for the entries. The search is made easier by drawing conclusions from the information given about the authors. However, this information about the genres produced by an author is often not correlated with specific works, so that one has to go through each entry of an author to determine whether it is a drama, a poem, a novel, an essay, and so on. Filtering options and more precise annotations would make research easier.
>
> Entries in the corpus of the Univerzitná Knižnicsa v Bratislave are easy to find thanks to good and precise search options. However, when it comes to Slovak novels, there are almost no results. If you search for novels, there are three results, and if you search for

"Slovak literature", there is only one result, which is not a novel. There must be something wrong with the structure of metadata and the metadata's relation to textual data.

The last corpus was the Slovak Literature Corpus. In my opinion, this corpus is more suitable for linguistic research on linguistic registers than on literature. There are search options for lemma, phrase and word form or character sequence, which is not helpful when looking for a specific literary genre and therefore not relevant for research in this case.

In summary, it can be said that improvements can be made in two areas.

Firstly, the quantity and quality of data. The representation of the Slovak novel is below average, so there is a need to expand the available data. The situation calls for investments in digitalization processes which should be conducted with a precise list of texts in mind. The researchers that are supposed to develop the list of texts have to explicitly answer to the above-mentioned problems of representativeness and balance in relation to Slovak narrative prose. As for the structure of the corpora: In order to make the data more available, the search options must also be made more intuitive and precise. It is very difficult to search for specific entries in large corpora if there are no options to narrow down the results. It can be seen that in the corpora above only one aspect is often fulfilled. On the one hand, the Zlatyfond SME 2.0 corpus has the largest amount of data available, but the search options are so poor that the search takes quite a long time. On the other hand, the DIKDA corpus, for example, has many filter options but few or no entries. So we need to improve both the search options and the content of the collections.

Of course, this requires an investment of resources, time, and money. This would make it easier for researchers in less affluent countries to participate more fully in international exchanges so that their experiences can be used by research and educational institutions.

Annotations and metadata become the main topic of our next case study.

## 3. The Modernist Haiku across European Text Collections

The refinement of data curation and selection is a fundamental part of Computational Literary Studies. We will use the exemplary case of scanning European text collections for instances of the genre "haiku" to present a prototype of the solution, i.e. the Corpus Exploration Platform, which, alongside other functionalities, will shortly communicate via APIs with other collections, thus initiating the environment of programmable corpora. On the face of it, the haiku, as a well-established and strictly prescribed short form, should be easy to identify in multi-genre and multi-lingual collections. However, difficulties arise from both cultural (Hokenson 2007; Johnson 2011; Śniecikowska 2016; 2021) and infrastructural factors.

### 3.1. Why Haiku? The Importance of the Genre and the Search

Literary history suggests that the haiku should be as ubiquitous as it is polymorphous in twentieth-century collections of texts. The genre is said to have had 'the most powerful poetic influence on Western poetry in the last century' (Jonson 2011: 6). The culture of modernism, especially symbolism and the avantgarde, appreciated the haiku's image-centeredness and non-narrativity. By virtue of these characteristics, haiku followed the path blazed by Japanese painting, without which the emergence of modern art would not have happened, or would have looked very different (Irvine 2014). The haiku, incidentally, appears in its original context in the visual genre called

*haiga*, typically painted by the haiku poets themselves (*haijin*). This opens up the prospect of multimedia approaches to CLS, which we unfortunately have to forgo in this case study. It is difficult to name a well-known modernist poet who has not been associated with the genre, which perhaps paradoxically makes the search for specific examples even more difficult. There are also significant chronological shifts within international modernism. While Anglo-American literature frenetically embraced the genre at the beginning of the twentieth century, it was not until 1975 that the genre really came into vogue in Poland (Śniecikowska 2021: 110).

These developments have produced an enormous variety of forms. Hoyt Long and Richard Jean So (2016), in a well-received article, searched for haiku in a pre-selected collection in English only. We are interested in multilingual collections and in the process of pre-selecting the texts themselves, i.e. identifying potential haiku candidates. The question of necessary features to identify arises: Long and So focus on "word-level features", using the bag of words technique, but also include in their representation "more complex formal features by recording their simple presence or absence in a text. Given the importance of syllable count to perceptions of the haiku during its early reception, we include this feature as well" (Long & So 2016: 255).

The multilingual environment makes our task more difficult, especially concerning the haiku's formal features.[11] Given the strict rules of Japanese phonetics (*mora*[12], *kireji*[13]), syntax and semantics (kigo[14]), the genre of haiku seems untranslatable into European literary systems. Where the haiku uses *kireji* (cutting words), the western poet typically divides the poem into verses. For example, "at the end of the first or second segments there may be the extra syllable *ya*, at most a gentle untranslatable 'ah,' indicating a rhythmic pause, or sometimes foreshadowing a change of theme or meaning. In contrast, the ending *kana* gives a sense of completeness, or 'Isn't it so?'" (Addiss 2012: 5). All European haikus are only approximations of the exact form, which is dependent on the specific structure of the Japanese language. This impossibility of identifying the canonical version of the haiku outside of Japanese literature (where the genre also evolves), coupled with the heterogeneity of the existing infrastructure of text collections, makes the search for sustainable data extremely time and resource consuming. The following questions arise: Which features of the different European approaches to the strictly Japanese genre should be given priority in identifying a poem as a haiku? How can the evaluation and exploration of literary text collections be facilitated in terms of these features?

### 3.2. Adapted Approach

To answer these questions, we selected several corpora for closer examination from the list of literary resources used also in the first case study.

Relying on metadata- and data-derived information, the corpora were scanned for both "explicit" and "implicit" ("adaptations" according to Long and So 2016) realizations. The liming factors were, of course, the time of publication and genre. Only those literary collections that contain twentieth-century texts and are tagged with the labels "lyric", "poem", "poetry" or have no tag at all should be considered. (The solution in section 3.3 potentially counteracts also the situation when a collection was falsely labelled.)

Explicit haikus are either translations of the Japanese models or original creations labeled as "haiku". Their findability relies heavily on metadata provided by corpus compilers or distributors. Two telling cases are Wikisource and the Russian National Corpus, which recognize

[11] The most extensive account on the formal features of the haiku remains Kenneth Yasuda [1957] 2001.

[12] A unit of duration in Japanese used to measure the length of words and utterances; mora languages differ from syllable languages like English.

[13] Cutting syllable, separating the first verse from the others.

[14] Codified signals of seasons: plus 1000 lexemes contained in the *saijki*, a prescriptive list of such words.

the haiku as a distinct form, but only in annotations to separate poems and not in the inventory of genres.

First, we embarked on manual search, which meant concentrating on explicit haikus since taking into account formal features would take an enormous amount of time. The fact that the corpora are set up differently makes it difficult, for example, to run a single Python script for all of them. Another problem has been the abandonment or re-launch of some sites, which often takes time. For example, the Korpus der Literarischen Moderne[15] (Corpus of Literary Modernism), which sounds promising for haiku, is currently unavailable due to a relaunch.

Nevertheless, some corpora, such as the Scottish Corpus for Text and Speech and Project Gutenberg, contain haikus, sometimes quoted in a work of a different nature, for example a description of Japanese sex life (the only result in Project Gutenberg). The text itself does not expand on the haikus, but the haikus illustrate some points. Of course, searching through long texts for scattered haikus costs time and energy. Table 2 contains the preliminary results:

| Corpus Name and Link | Metadata | What is the corpus about |
|---|---|---|
| Europeana:<br>https://www.europeana.eu/de<br><br>1)<br>https://www.europeana.eu/de/item/92006/BibliographicResource_1000095231552<br><br>2)<br>https://www.europeana.eu/de/item/92006/BibliographicResource_1000095231588<br><br>3)<br>https://www.europeana.eu/de/item/92006/BibliographicResource_1000095231777<br><br>4)<br>https://www.europeana.eu/de/item/2023859/08400_08419_borito_jpg | 1) Haiku in the luggage;<br>Terebess, Gábor<br><br>2) Haiku a poggyászban,<br>Terebess, Gábor<br><br>3)Szigetláko;<br>Kelemen,<br>Hunor<br><br>4) Haiku;<br>Szenti, Tibor (author)<br><br>5) 44 haiku;<br>Terebess Kiadó [Publication]<br>Terebess, Gábor (translator) Rjókan,<br>Taigu (author)<br><br>6) Modern nyugati haiku: Terebess Gábor fordításai;<br>Terebess, Gábor (translator) Steinert,<br>Ágota (editor)<br><br>7) Százhetven haiku;<br>Terebess Kiadó [Publication] Fodor,<br>Ákos (translator) Racskó, Ferenc | 3) 4 Haikus |

[15] https://kolimo.uni-goettingen.de/index.html

| | | |
|---|---|---|
| 5)<br>https://www.europeana.eu/de/item/2023859/03700_03795_borito_jpg<br><br>6)<br>https://www.europeana.eu/de/item/2023859/06200_06241_borito_jpg<br><br>7)<br>https://www.europeana.eu/de/item/2023859/03800_03832_borito_jpg<br><br>8)<br>https://www.europeana.eu/de/item/174/dia_Kanyadi_Sandor_Valaki_jar_a_fak_hegyen_kanyadi00297<br><br>9)<br>https://www.europeana.eu/de/item/9200479/item_48239065_ | (selected by) Racskó, Ferenc<br>(translator) Matsuo, Basho (author)<br><br>8)<br>Három haiku haiku témára;<br>Kantor Peter Sándor Kányádi<br><br>9) Haiku i literatura polska przełomu XIX i XX w. : o estetycznej wartości nastroju;<br>Szymańska, Beata | |
| Scottish Corpus of Texts & Speech:<br>https://www.scottishcorpus.ac.uk/<br>1)<br>https://www.scottishcorpus.ac.uk/document/?documentid=922<br><br>2)<br>https://www.scottishcorpus.ac.uk/document/?documentid=129&highlight=haiku<br><br>3)<br>https://www.scottishcorpus.ac.uk/document/?documentid=130&highlight=haiku | 1) 50 Haikus by Japanese Masters;<br>Purves, David<br>2) Scots Haikus;<br>Leeming, Mr Bruce<br>3)<br>Scots Haikus II;<br>Leeming,<br>Mr Bruce | Rather about Haikus and Scottish translations |

| | | |
|---|---|---|
| Projekt Gutenberg – DE<br>https://www.projekt-gutenberg.org/<br><br>1)<br>https://www.projekt-gutenberg.org/krauss/japgesch/chap003.html | 1)<br>japanisches Geschlechts-leben,<br>Friedrich S. Krauss | rather a description of Japanese culture with sometimes haikus in between |
| Biblioteka Maksima Moshkova<br>http://lib.ru/<br><br>1)<br>http://lib.ru/POEZIQ/GrepSearch?Search=Haiku | 1)<br>haikus,<br>links,<br>anthologies / collections,<br>texts about haikus and translations | Corpus links to other websites collecting haikus |
| Hathi Trust Digital Library<br>https://www.hathitrust.org/<br><br>1)<br>https://catalog.hathitrust.org/Search/Home?lookfor=haiku&searchtype=subject&ft=ft&setft=true | 1)<br>Japanese haikus and texts about haikus and Japanese haikus | No haikus out of Japanese context as it seems |
| Biblioteca Virtual Miguel de Cervantes<br>https://www.cervantesvirtual.com/<br><br>1)<br>https://www.cervantesvirtual.com/buscador/?q=haiku | 1) mostly texts about haikus | No search for genre etc |
| World Digital Library<br>https://www.loc.gov/collections/world-digital-library/about-this-collection/<br><br>1)<br>https://www.loc.gov/collections/world-digital-library/?q=haiku | 1)<br>Japanese books and original prints | |
| Federacja Bibliotek Cyfrowych<br>https://fbc.pionier.net.pl/<br><br>1)<br>https://fbc.pionier.net.pl/search#fq={!tag=dcterms_accessRights}dcterms_ | 1) some Japanese prints and books about haikus | |

| | | |
|---|---|---|
| accessRights%3A%22Dost%C4%99p%20otwarty%22&q=haiku | | |
| Open Texts World https://opentexts.world/ | Many results, when searching for haiku: books about haikus and haikus | |

(Table 2: Results of our manual search for haikus)

As for implicit haikus, their findability depends on the addressability of certain textual features (such as: verses, syllables, and topics); the addressability results from the structure of a corpus. In some corpora there are advanced search options. For example, in the Diachronic Spanish Sonnet Corpus, one can search for rhyming words and see graphs of rhyme frequency, as well as stress and enjambment. However, the corpus contains only sonnets.

SPARQL queries in the LINHD POSTDATA project provide a clear overview and uniformity of metadata that is lacking in most corpora.

As for implicit haiku, whose findability depends on the accessibility to the formal features of texts contained in a collection: JSON works well as a format for presenting ordered metadata, and we have made use of it. As a result of favouring the cultural universal of verse number over the relation to syllable number (absent in Japanese that uses mora), an algorithm was developed to extract three-verse poems from the DLK (Deutsches Lyrik Corpus, cf. Haider 2021). In the DLK, verses and syllables are already marked up in JSON, which makes it possible to count syllables and thus provides an additional criterion for identifying possible haikus. However, the ideal structure of 17 (5+7+5) syllables could not be identified.

There are two problems: inconsistencies in the recording of metadata and discrepancies in the structural annotation of texts in literary corpora.

### 3.3. Proposed Solutions: Corpus Exploration Platform

Disclaimer: This section represents the provisional state of a work in progress, the final form and appearance of which may differ significantly from what is presented here at this time.

We believe that the above problems predestine modernist haiku to serve as a benchmark for the implementation concepts currently being developed within the CLS INFRA project. Our rationalisation scheme will be successful if we are able to locate a larger number of haiku in multilingual literary text collections and help other researchers with similar tasks.

Specifically, a prototype of the Corpus Exploration Platform (CEP tool), whose main function is to catalogue literary textual resources, will be constructed as an instance of the open-source knowledge management system ResearchSpace (researchspace.org), which the British Museum has made available to numerous digital humanities projects.

ResearchSpace, and by extension the CEP, aims to (re)create dynamic relationships between objects and events. These are said to go beyond the tabular forms of relational databases and to be semantic in nature. Like any other project implemented with ResearchSpace, CEP uses an ontology as a framework for representing knowledge in the form of structured but multidimensional information. The ontology is stored using a graph database containing RDF-encoded representations of corpora and their documents as Linked Open Data, and equipped with a customisable set of patterns for linking such data to external reference resources. Such a

dynamic knowledge graph is stored as triples and can be manifested in various data interactions, visualisations, rationalisations and workflows using ResearchSpace's templating system for front-end development. According to the creators of ResearchSpace, this way of relating data to each other reflects both the way the human mind works and the intricate reality; we definitely acknowledge the complexity or even convolution of the data landscape.

The ontology for CEP, as a concrete instance of ResearchSpace, is conceived as a descriptive metamodel for literary corpora, providing controlled vocabularies for categorising the corpora, documents, and files, thus allowing a fine-grained genre-specific description of collections, including a list of formats, methods, tools and features.

The formal ontology conforms to the CIDOC CRM, while building on the "Metamodel for Corpus Metadata" (MCM; Odebrecht 2018); this model, which has been proven to work well for linguistic corpora (cf. Laudatio Repository), should be rendered in a machine-readable format. Furthermore, it needs to be adapted to literary text collections and queries that are made with literary research questions in mind. This process of adaptation amounts to extending the CIDOC CRM with reference to established ontologies, such as FRBRoo/LRMoo or Postdata's Ontopoetry Core Ontology, as well as reference resources, such as Wikidata or the Dewey Decimal Classification. We have specified the latter with a more detailed list of literary genres and devices.

The CEP will consist of several components, but its core is a catalogue of literary corpora which, in addition to the list of resources (including a special class of "top level corpora"), contains explicit statements about the various characteristics of individual corpora - in particular information about the internal structure of the documents, e.g. the presence of explicitly annotated verses. By recording provenance information, CEP will allow even conflicting statements about a resource to be made, which can be traced and evaluated individually.

Data formats are also listed in the catalogue. This is very important for the architecture of the CEP, because another component of the CEP is a catalogue of research tools and methods, mapping formats to tools or methods. Transformation paths between the most common formats and encodings are described in a so-called transformation matrix.

Another component of the CEP is a SPARQL endpoint that allows exploration of the catalogue by relevant categories and relationships between them, while browsing the knowledge graph. A SPARQL API allows communication with other resources stored in RDF databases, such as the POSTDATA (Poetry Standardization and Linked Open Data) project or DraCor (Drama Corpora Project). The triangulation between CEP, POSTDATA and DraCor, communicating via APIs, is a first bud for the future environment of programmable corpora, defined in CLS INFRA as "corpora that provide an open, transparently documented and (at least partially) research-driven API to make texts machine-processable".

In the haiku example, this approach will be tested by attaching information such as tokenisation rules, syllable structure, stanza structure, etc. to the corpora according to the features described in the previous subsections.

## 4. Measuring Entropy and Surprisal in the Prose of the Tsarist Empire Devoted to Terrorism (Russian and Polish Texts)

The development of the political strategy of terrorism in the Tsarist Empire went hand in hand with or was preceded by, the creation of a large number of narrative texts.

The corpus of these narratives of terror, conspiracy, and provocation contains many significant works: alongside Belyi, we have Dostoevsky‘s *Demons*, narratives by Tolstoy, Chekhov, Andreev, Brzozowski or – in a global scale – Joseph Conrad. But the terrorists’

proclamations also use narration, fabulate, and relativize their reference to reality with regard to their effect (they provoke facts rather than state them while oftentimes retaining the form of report).

Those texts, both belletristic and belligerent, are endowed with specific characteristics due to their thematic and formal orientation toward the violent activities; the description of these characteristics firstly enables a cultural-historical reinterpretation of the evolution of the fictional discourse of the Tsarist Empire between 1860 and 1914 and secondly provides new insights into the narrative construction or reproduction of consciousness (percepts, cf. Schmid 2017). Thirdly, this historical narrative mode influences how terrorism is conceptualized and experienced today (for example in the recent case of the organization claiming responsibility for Daria Dugina's death or Twitter gossips that Putin`s carriage, pardon, limousine was pelted with bombs). Fourthly, this mode brings to light the workings of economy in narrative texts, economy here meaning not only the actions involved in producing, exchanging, and consuming goods and services, but also the budgeting of energy to achieve aesthetic and social efficacy.

### 4.1. Literary Historical and Theoretical Assumptions

The fact that the poetological or narratological question reciprocates the ideo-historical and economical ones is already evident in the title's term 'terrorist realism' (henceforth TR), which draws on Roman Jakobson's conception of 'revolutionary realism in literature'. Jakobson describes 'revolutionary realism' as a kind of estrangement consisting in an abrupt reversal of semantic hierarchies (Jakobson [1921] 1969: 379). With this formulation, Jakobson relates what has become the established concept of estrangement (*ostranienie*), which justifies or explains literary evolution, specifically to narrative, social upheavals, and hypothesizing about the make-up of the outward world, upweighting or downweighing the significance of particular qualia for the assessment of a situation so as to produce the reality by way of hypostatizing your hypotheses (akin to today's 'predictive processing', cf. Friston & Stephan 2007). By adapting and shifting Jakobson's term, we claim that it refers not to 'revolution' in all its ambiguity, but specifically to 'the individual terrorism' of *sztyletnicy* (dagger fighters), *Narodnaia volia* (the People's Will, or Freedom), the party of the Socialist Revolutionaries, or the Polish Socialist Party—to what is known in the West as the 'Russian method' (Patyk 2018: 230). we primarily seek to avoid confusion with the Marxian concept of 'revolution' as an expression of the mature class consciousness of the proletariat, able to recognize its economic interests: The difference to Marxism lies at the heart of the most striking characteristics of TR—instead of the scientific-objective view of historical processes, which supposedly characterizes the Marxism of the Second International, TR focuses on the theme and technique of provocation, using violence and fabrication to force the transition to the desired state. The reader should experience not maturity, but rather its antithesis—acceleration, prematureness, impatience, and insecurity. It is made conspicuous through the dynamization of the narrative perspective, which foregrounds shifting attention and hectic hypothesizing, i.e. re-hierarchization and reinterpretation of the percepts. The poetics thus refers—affirmatively or disapprovingly—to concrete political and ideological positions, particularly.

In my attempt to translate populist economics into aesthetic and narrative conventions, I will draw on the work of the sociologist Leon Winiarski from the 1890s (1967), who was forced to flee the Tsarist Empire in 1885 as a result of the persecution of the first Polish workers' party, the Proletariat. Despite its name, the party – as Rosa Luxemburg (1897) pointed out – had a Blanquist, populist disposition, as it did not assist in the prenatal development of proletarian class consciousness, but formed conspiracies and even an alliance with the Russian terrorist organisation The People's Will. As a *privatdozent* in Geneva, Winiarski worked on extending the

economic theories of the Lausanne School, sometimes called the Mathematical School (Léon Walras, Vilfredo Pareto), to society as a whole, with particular attention to aesthetics. In lucid French prose, backed up by mathematical formulae, Winiarski conveys, alongside or perhaps through the general tenets of the Lausanne School, views that correspond to populist terrorism, such as the recognition of individual genius as the only factor of social change, as opposed to society, which is conservative by default. He also expresses a view, typical of both terror and modernist art theory, that the utility, the enjoyment, of a state is directly proportional to the effort required to achieve it, as if Winiarski could not fully accept the Lausanne School's doctrine that the marginal utility tends to diminish after passing a threshold - too much of a good think may be taxing. Occasionally Winiarski seems to agree with Mae West's quip "Too much of a good thing can be wonderful!".

Winiarski expresses this invitation to or justification of the energetic expenses in equations derived from Lagrangian/Hamiltonian mechanics. These equations reveal an economic or energetic rationality behind the anti-Marxist Eastern European left. Moreover, these equations make Winiarski's theory compatible with the current theory of active inference as a way of understanding sentient behaviour. According to the Lausanne School, the ultimate source of value is not labour or the exploration of labour, as Marx would have it, but utility or rather a belief in utility. Accordingly, Winiarski argues that it is not economics but the transformation of energy that provides the most general explanation of both physical and social realities. Energy, by definition, changes form, so that the aesthetic energy of enjoying beauty can be transformed into the political action of throwing bombs, or vice versa.

It is precisely the conspicuous qualities of TR, the form of subjectivity with its temporality and spatiality, which relate to its rootedness in intellectual and political history, full of violence and provocation, that are most relevant for its operationalisation, i.e. translating these qualities into computable quantities. This is done with the help of the concepts of surprise and Kullback-Leibler (KL) divergence. According to the proponents of active inference, these measures correspond to Hamiltonian mechanics (see section 4.2).

The model of TR that we propose combines the classical communicationist approach with postclassical cognitivism (specifically enactivism, predictive processing, and active inference, all of which I am subsuming under the umbrella term 'inferential enactivism') to provide a framework for characterizing the complicated formal relationship between terror and literature. Acts of violence and utterances can be aligned with one another at a more general level of communication, since both are to be understood as acts of communication. *The Revised Academic Definition of Terrorism* describes acts of terrorism as "threat-based communication" (Schmid 2012). The narratological model hinges on the cultural-historical thesis that terrorists and their opponents (the secret police) generally communicate through provocation, which includes both acts of violence and counterfactual statements; moreover, the act of violence, understood as a provocation, shares with the fictional discourse specific traits that are their exclusive preserve. The authors of TR explore precisely these disturbing similarities between the polar opposite forms of communication, either in order to amplify the cause of the terrorists or to counteract it. For the sake of the affect (threat, terror) and mindfulness (awareness) that terrorist communication mobilizes, classic 'communicationism' in narratology (Sławiński 1971; Bartoszyński 1971) is supplemented with cognitivist elements. Consistent with the prevailing view (Olson 2011), We position 'classical' (structuralist) narratology as a kind of base upon which newer, 'postclassical' approaches to narrative can build.

Cognitive narratology has never been applied to the literary material that makes up the corpus of TR, although this is a particularly bodily-centered, affectively and cognitively charged form of communication. The same holds true for computational literary studies.

As TR foregrounds provocation and permanent ambiguity and even fear-induced paranoia, of all cognitive approaches, predictive processing and active inference[16] are arguably ones of the most suited. We will refer to my narratological extract of those biophysical and neurological theories as inferential enactivism. Inferential enactivism construes the subject's cognitive acts as world-making taking place in body-related and emotionally charged exchanges with the environment. And thanks to the predictive nature of its worldmaking, the study of it can be based on a quantitative analysis of the TR text corpus, especially with a view to informational entropy and surprisal.

TR, in its capacity as literature in the second degree, models within its safe "as if" mode the relationship between the counterfactual and the factual, a distinction which provocation upsets to the point of delirium.

Dostoevskii, Brzozowski, Belyi, and Conrad construe their provocateurs as fabulists restlessly shifting the limits of the factual by self-fashioning and staging simulacra of reality. The actual terrorists – the models for the writers – penned proclamations designed to blur what the case is.

The implied basis of comparison between factual violence and narratives is communication. The narratological point of reference is furnished by the Polish school of literary communication ('Warsaw Structuralism', cf. Sławiński 1971; Bartoszyński 1971, Balcerzan 1971), which draws on Viktor Vinogradov (1930), Valentin Vološinov ([1928] 1985), and Roman Jakobson ([1960] 1987). According to the United Nations Consensus Scientific Definition (Schmid 2012), acts of terrorism are acts of communication that can be characterized by the presence of a division into "direct targets" and "main targets". In violent communication, direct targets serve only as the generators of messages addressed to the main targets (dead politicians cannot be scared into making concessions). As a rule, in the context of the Tsarist Empire, this split in the receiver entity corresponds to the split in the emitter entity, divided, in the spirit of provocation and conspiracy, into an authentic and a counterfactual sender. In order to generate revolutionary energy, to maintain conspiracy and to increase uncertainty, the actual, historical terrorists created an imaginary sender: the Revolutionary Central Committee (as the sender of the first terrorist manifesto Molodaia Rossiia, penned by the imprisoned student Zaichneveskii), Bakunin and Nechaev's Narodnaia rasprava, etc. In fictional narratives, the terrorist groups are usually presented as revolving around at least one strong manipulative character who fulfills the role of an author affecting the borders of the factual. So the groups are necessarily split into narrators (emitters) and audience, whose members appear as actors for another audience, both inside and outside the diegetic world.

It is precisely this split in the communicative entities that allows the act of terror to be set side by side with the literary work, which always has an implicit and an explicit reader or author too. According to Balcerzan (1971), literary narrative texts are the only type of utterances that necessarily represent the process of sending and receiving alongside the diegetic world, i.e. they obligatorily divide their entities into inner- and extratextual, fictional and factual. Likewise, according to Foucault, only *authored* texts show a discontinuity between the fictional speaker and the empirical writer; the 'author' function, which is fundamental to fiction, operates precisely in this hiatus (Foucault [1970] 2014: 22-24).

If the division of communicative instances and the reflexive representation of this division is the most fundamental characteristic of literature, then literariness overlaps with the social

[16] Predictive processing has recently become widespread in cognitive science and neuroscience (see Friston 2012; Seth 2013), philosophy (see Hohwy 2013; Clark 2015) and narratology (Kukkonen 2014, 2016, 2017, 2018 & 2019; Kukkonen et al. 2014; Caracciolo 2014, 2016; Popova 2015; Irving 2016).

function of literature, that of communication, whatever the aim of a work. An overflow of aesthetic energy into political spheres, as postulated by Winiarski, does not seem so implausible from a narratological perspective.

Provocation – the main topic of Dostoyevsky, Conrad, and Belyi's novels – pertains not only to manipulative utterances aimed at gaining actual advantage, but also to the actual killing, which affects the mental world models and compels adversaries to respond ("Zugzwang").

The actions of the terrorists and their adversaries thus inscribe themselves in the basic situation of communication, which consists in exchange. An action on the part of the sender is supposed to provoke a reaction so that the receiver in turn adopts the position of the emitter. For example, the desired reaction to an attentat – as stated in the first terrorist manifesto "Molodaia Rossiia" – is a wave of reprisals that arouse hostility between the people and the government and thus raise class consciousness.[17]

Inferential enactivism explains the origin of energy and its transformations in the communicative process. The theory introduces to the mix the active body without which energy lacks a scientific meaning. The body is described in terms of a crack Kantianism: perceptions and other cognitive processes arise in the embodied mind as a result of the incessant activity of this body; perceptions and emotions are solicited and not simply had.

The starting point of enactivism is the assumption that cognition and consciousness have a primarily life-sustaining function (Seth 2021: 7); as with Winiarski, in active inference these exchanges and transformations of energy are expressed in the equations of Hamiltonian mechanics, which are seen as equivalent to conservation of energy and probability (informativeness); probability related to Bayes theorem, form the subjective point of view, and to surprisal and entropy, as view objectively.

The brain is generally assigned the role of manager of energetic resources; for the sake of economy, the brain opts in a predictable environment for an automatism of reactions, so that awareness amounts to a state of exception that arises in moments of threatening uncertainty as the need for decision-making arises.

Inferential enactivism follows Helmholtz' understanding of the relationship between the brain and perception as the result of the former's 'unconscious inference' (*unbewusster Schluss*, 1867, 3rd volume; Friston & Stephan 2007). The subject perceives its own projections as appearances of the external and internal worlds. The brain enclosed within the cranium 'infers' from the electrical stimuli reaching it by relying on previous similar situations what their cause might be and what it should most likely prepare the body for. The world turns out to be a sum of imagined or assumed causes of sensations coming from outside and inside the body (the inner and the outer world as the hypostasis of probabilistic hypotheses). The subject is immersed in their own hypotheses as to the true nature underlying sensations.

To remain in the organism's preferred range of states, the brain tries to minimize 'suprisal' or prediction errors. While perceptual predictions flow predominantly 'top down' (Friston 2005; Solms 2021: 142), prediction errors flow 'bottom up': the brain only conveys upwards the portion of the incoming information that does not meet expectations (Hohwy 2013, Clark 2015: 23). These prediction error signals disrupt habitually-based automatism and trigger awareness; the brain then uses the prediction errors to update its predictions.

Already Helmholtz related predictions based on earlier examples to Bayes' theorem. For literature, especially modernist prose, prudent (anticipatory) recourse to past experiences

[17] The terrorist tactics suits particularly well the description in terms of information theory equating entropy with informativeness (Shannon 1946): the bomb is nothing else than a message *sub specie* the discharge of energy. Now, that type of information must be positioned as a compound that cannot be reduced to the intelligible or the noetic by virtue of the information's inalienable affective and sensorimotor components – the bombastic message is directed at the body.

becomes relevant in that literary utterances bring together the accustomed and the unexpected. The accustomed serves as a kind of background against which a character becomes consciously perceptible by differing from backdrop as a prediction error (startling juxtapositions of words or 'events' in narratives). If phenomena cross the threshold of 'mechanical' recognition, the recognition of patterns gives way to awareness, to 'seeing'. Estrangement, which is fundamental to the TR, can be understood in the sense of predictive processing as an artistically shaped series of prediction errors. The aesthetics of modernism and provocation make subjects out of the ordinary; they seek to maximize surprisal. It is noteworthy that hitherto no one has linked predictive processing and estrangement, let alone TR.

Estrangement, which is fundamental to the TR, and eventfulness can be understood as artistically shaped series of prediction errors.

It is noteworthy that no one has yet linked predictive processing and alienation, let alone TR. Conversely, linguistics measures the two types of surprise that active inference addresses: "how poorly a model fits the data it is trying to explain" (the appearance of something seems unlikely) and "a measure of how much we have to update our beliefs following an observation" (Parr et al. 2022: 18, 20). Both formalisms - Surprise and KL Difference – show difference between different language registers, which means that style can be inferred from the management of surprise in relation to a communicative goal.

Inferential enactivism docks onto classical narratology in at least two substantial ways. First, it helps to explain and operationalize the notions of estrangement and eventfulness. Second, it provides a new energetic view of communication that computationally links factual terror with fictional and non-fictional literature as two forms of informational exchange beyond dull parsimony. Here, an excess is rational, just as awareness beats automatic responses.

The prominent representative of classical narratology Jurii Lotman states that without surprisal there cannot be eventfulness and significance in narrative. As is well known, Lotman characterizes the event and the plot – i.e. a chain of events – as always transgressive. Compliance with the norm is uneventful (Lotman [1971] 1977: 231 ff). While translating Shklovskii's theory of estrangement into the language of structural semiotics, Lotman correlates the norm with expectation and event with surprisal:

> A device is viewed in terms of the general conception of [slow perception,] and the de-automatization of form as the relation between expectation and the text. Thus for Sklovskij a device is the relation of an element of one syntagmatic structure to that of another, and consequently it includes a semantic element (Lotman [1971] 1977: 232).

In the parlance of enactivism, one explanatory model replaces the previous one that fails to account for a strange phenomenon. This is the translation or transition of the subject from one explanatory framework into another, a gestalt switch, that produces both significance and eventfulness. "An event in a text is the shifting of a person across the borders of a semantic field" (233). (As Michel De Certeau put it: "What the map cuts up the story cuts across" (De Certeau 1984: 129)). It is in connection with this that Lotman calls the event "a revolutionary element opposing accepted classification" (Lotman [1971] 1977: 234) and the plot "the 'revolutionary element' in relation to the world picture" (Lotman [1971] 1977: 238).

In terms of communication theory, prediction also applies to our relationships with other people, including those who wish to provoke or kill us. In relation to "mutual prediction", Andy Clark speaks of "self-fulfilling psychosocial knots and tangles" (Clark 2015: 76), which sums up much of TR.

Rainer Paris's communication-based theory of provocation, apparently modelled on individual terrorism, involves the adaptation of a model of reality to account for a prediction error. A model of reality based on habit is updated when the provocation "elicits a reaction that, in the eyes of the third party, morally discredits and unmasks [the provoked opponent, i.e. the indirect target]."[18] Discreditation entails a vehement adaptation of a model in view of a flagrant prediction error. In literary texts, as communicationism and structuralism tell us, the third party is located both inside and outside the text; both instances are caught up in a transformation of energy. But provocative violence begins with the manipulation of reality in a maximally unpredictable way; provocation amounts to "a deliberately committed but unexpected violation of norms intended to draw an opponent into open conflict".

The effectiveness of violent provocation depends on the degree of blatancy of the initial violation; the extent to which the provocateurs enforce the updating of beliefs depends on how much they upset the balance of the social system. This is the logic of the economy of terror, which is also present in modernist art: the greater the surprise and the investment of energy on the part of the sender, the more effective the provocation is in relation to the recipient. Moreover, this paradoxical economy of effort is at the basis of our existence as living, feeling and cognitive organisms, as demonstrated by the mathematical formalisms used by Winiarski and active inference.

If we express the communicative situation of terror or TR in terms of energy transfers, the initial provocation, the opening of the exchange, appears as a great influx of energy; the greater the influx, the greater the surprise. Winiarski reminds us that, according to Lagrangian mechanics, every disturbed system tries to regain equilibrium. However, contrary to unscientific intuition, it is wrong to associate equilibrium with boredom. Active inference also warns us that the organism's quest to minimise free energy "should not be equated with a rigid repertoire of responses (e.g. automatic homeostatic responses), but rather the opposite, especially in advanced organisms. We can develop open-ended repertoires of novel behaviours to pursue our original homeostatic imperatives" (Parr et al. 2022: 60). According to Winiarski, equilibrium means the most energetically possible state of the system:

> Un système économique sera dit en equilibre lorsque, après avoir été dérangé de la position qu'il occupe, on a fait par là même naître dans le système des forces qui tendent à le ramener à la position primitive. […] Le principe suprême de la Dynamique, celui de Lagrange, nous dit que chaque système matériel tend à produire, au moment de l'établissement de l'équilibre, le maximum d'énergie compatible avec les conditions existantes. (Winarski 1967: 95, 168)

The only thing that prevents the economic system from achieving the maximum amount of energy are monopoles (Winarski 1967: 95); probably analogous inhibitions apply to the field of art. According to active inference, these are built-in limitations or preferences regarding the environment that prevent organisms from turning into excess machines in terms of energy and information. However, "if one removes both ambiguity and prior preferences, the only remaining imperative is to maximize the entropy of observations (or states […]). This may be interpreted as uncertainty sampling (or keeping one's options open)" (Parr et al. 2022: 36). The hunger for new information, whatever the cost, is the most primal drive of living organisms from a mathematical point of view.

[18] We are using here Lynn Patyk's translation (Patyk 2023: 11). Orig: Paris 2015: 49: „Ich definiere eine Provokation als einen absichtlich herbeigeführten überraschenden Normbruch, der den anderen in einen offenen Konflikt hineinziehen und zu einer Reaktion veranlassen soll, die ihn, zumal in den Augen Dritter, moralisch diskreditiert und entlarvt., Der Wille des Einen ist das Tun des anderen."

It follows from both Winiarski's and active inference's equations that jouissance, pleasure, utility and expenditure of energy (X, Y, Z) must be proportional to effort or sacrifice.

$$\Sigma \left[ \left( X - m \frac{d^2x}{dt^2} \right) \delta x + \left( Y - m \frac{d^2y}{dt^2} \right) \delta y + \left( Z - m \frac{d^2z}{dt^2} \right) \delta z \right] = 0.$$

$$F[Q,y] = \underbrace{-\mathbb{E}_{Q(x)}[\ln P(y,x)]}_{Energy} - \underbrace{H[Q(x)]}_{Entropy}$$

Fig. 1. Measures of effort and entropy should be as big as possible if free energy F[Q,y] ist to be as small as possible (Winiarski 1967: 169; Parr et al. 2022: 28)

The entropy in active reasoning or effort (biological energy transformed by an individual) in Winiarski must be large enough to balance energy and pleasure, respectively. Thus, under certain circumstances, it turns out that minimising free energy means striving for maximum entropy (randomness, informativness). “The agent is striving to *increase* entropy [...] of the agent’s (approximate posterior) belief. […] [T]he imperative to explain things as accurately as possible but also “keep options open” and avoid committing to any specific explanation unless this is necessary – that is, the *maximum entropy* principle (Jaynes 1957)” (Parr et al. 2022: 60). The agent thus invests energy in maintaining their own uncertainty. In terms of aesthetic equilibrium, Winiarski explicitly says that the intensity of pleasure is proportional to the expenditure of energy or money:

> La loi fondamentale de l’économie mathématique, que peut du reste être déduité des équations générales du mouvement de Lagrange, est la suivante: l’équilibre dans la satisfaction des besoins ne s’établit que quand l’intensité des derniers désirs satisfaits est proprotionelle aux dépenses en énergie (ou en argent) (Winarski 1967: 198).

The energetic expenditure – at least in its aesthetic form, not necessarily as an attentat – brings indispensable adaptive advantages, especially for top-down modelling, as the organism can train itself to formulate and test hypotheses when confronted with puzzling patterns. Proponents of enactive inference emphasise the indispensable value of the ability to formulate counterfactual reasonings for the enactment of the factual. However, it is worthwhile for a counterfactual model to have aesthetic features as well (Clark 2017: 277)

TR's texts contain instructions for formulating and testing models, with gestalt switches triggering eventfulness and aesthetic pleasure. "In narratological enactivism, the revisions of prediction throughout the narrative provide the outline of a plot trajectory" (Kukkonen 2019: 20). And the texts themselves, as aesthetic artefacts, tend to generate prediction errors. TR both models the cognitive processes of its communicative instances and directs the process of reading; in the second case, TR makes the reader a participant in the communicative process, i.e. a part of the model.

## 4.2. Corpus Compilation for Computational Research

We propose the following expression of the narratological model in terms of computational linguistics[19] and thus computational literary studies:

The common ground for corpus linguistics and the narrative model, which includes enactive inference, is provided by the adaptation of energetic models in information theory, in particular the notions of entropy and its correlate surprisal. In linguistics, these notions of statistical mechanics that became a part of information theory – entropy and surprisal – are correlated with the effort the subject invests in encoding messages. Winnicki's reduction of aesthetics to Lagrangian mechanics, correlated with the measurement of effort and pleasure, allows us to specify the linguistic assumptions so that they apply to aesthetic forms of communication, such as artistic prose or the novel.

TR combines the biophysical concern with the survival of the organism as a separate entity in an environment with the linguistic parsing of signs, which as potentially ominous are of course related to life-threatening situations. The interpretation of signs explicitly assumes a survival function and is placed in an energetically charged model of communication.

In linguistics, effort (and entropy or surprisal) correlates with the more general problem of how human speakers with their finite resources can produce the potentially infinite richness of linguistic phenomena. The focus, as in TR, is on the finite character of memory and thus on a restricted range of predictions (world and discourse models), resulting in surprisal.

The finiteness of the subject corresponds in linguistics to the Uniform Information Density Hypothesis, which states that language production and reception show a preference for uniform distribution of information. The increased information density (<> surprisal or relative entropy) causes a slowing down of perception so that the perceiver's effort remains at a constant level. This means that the speaker can only invest a certain amount of energy in parsing an utterance. Memory and/or expectation explain the uniform information density because they are the instances that tend to sense effort and accordingly limit (slow down) their activities. In terms of energetic measures (effort), the difficulty of parsing a word w could be taken as the magnitude of the shift in the allocation of finite resources.

The narratological model of TR can be operationalised on the basis of two hypotheses:

Hypothesis 1: The narratives that belong to the class TR foreground the interplay of memory and surprise, therefore tend to expose a higher than average linguistic measures of entropy and surprisal; there is a correspondence between the subject matter and the linguistic features of the texts.

Hypothesis 2: If the TR represents a stable narrative mode, then its development over time should be characterised on the linguistic level by the maintenance of a high degree of relative entropy (related to the unstable, dangerous composition of the diegetic world) and a decreasing degree of surprise (the conventionalisation of a genre). Like the detective story, TR should combine a high degree of conventionality in its construction with the foregrounding of uncertainty at the level of the diegetic world. As will soon become clear, this hypothesis entails a strict distinction between the notions of entropy and surprisal.

[19] We relied mostly on the following works: Brouwer et al. 2021; Degaetano-Ortlieb 2019; Degaetano-Ortlieb & Piper 2019; Degaetano-Ortlieb & Teich 2022; Levy 2015; Lowder et al. 2018; Noortje et al. 2019; Pichler & Reiter 2022; Sayeed et al. 2015; Venhuizen et al. 2019; Venhuizen 2019a; Venhuizen et al. 2022.

B. Hypotheses Testing
The testing of the hypotheses begins with the construction of a reference corpus containing the most representative realisations of the narrative mode TR, identified on the basis of a critical evaluation of the scientific literature on the themes of realism and terror in the literature of the 19th century. The reference corpus is confronted with the archive (the corpus of the nineteenth-century novel, or perhaps the reference corpora of a given language in a given period).

The construction of the corpus has a threefold objective:

1. To train with help of the reference corpus a classifier that would find in the archive of 19th century prose similar texts that are not necessarily identified as belonging to realistic prose devoted to terror, including cases that tend to escape the perspicacity of human readers, and thus measure the impact of the mode on the literary/discursive cosmos.

In this module, a stylistic and thematic relationship between fictional and non-fictional prose texts devoted to terror can be statistically evaluated.

A tension that is said to be constitutive of artistic prose (Efraimova 2022) can be tested, namely between the artistic and the everyday use of language. This tension relates to the particular political engagement of prosaic *écriture*.

2. To study the language of the narrative mode of TR, as represented by the reference corpus (potentially enriched by the results obtained by the classifier), on the basis of two information-theoretic measures:
Kullback-Leibler divergence (KLD, called “relative entropy”) and surprisal (in a perfect world, the study should juxtapose the results obtained on the basis of n-grams and the measurement of “semantic surprisal”).

2.1. To measure the diachronic differentiation between the reference corpus and the archive(s): TR should look increasingly different from the archive over time, thus consolidating itself as a distinct narrative mode.
This process may include measuring its distance from different novelistic genres.
Measure: KLD

2.2. To test whether phrasal (i.e. lexico-grammatical) standardisation/conventionalisation occurs, which implies that surprise should decrease over time.
Measure: First, we will use KLD to identify the most characteristic/relevant linguistic features involved in diachronic change and then measure their predictability/information content with surprisal.

3.To measure, within particular texts, online entropy and surprisal, correlated with the process of constructing diegetic worlds and reading, respectively. The interplay of linguistic knowledge and world knowledge in the narrative texts amounts to a combination of linguistic and situational surprisal on the one hand with situational entropy, on the other, as the words produce large semantic figures (the elements of the worlds presented that relate in some way to the reader's world knowledge).
Measure: a multidimensional vector space in which the distances between vectors are calculated online using the cosine distance between words.

In the vector space, the distinction between relative entropy and surprisal, which are sometimes confused, becomes more pronounced; this in turn illustrates the paralogisms of the modernist theory of surprise or strangeness as the definiens of (literary) art.

As for the difference, linguistics distinguishes between state-by-state expectation (surprise) and end-state confirmation (entropy reduction). This corresponds to the two types of surprisal that active inference takes into account: “a measure of how much we need to update our beliefs after an observation” and “how poorly a model fits the data it is trying to explain” (the appearance of something seems unlikely) (Parr et al. 2022: 18, 20). Here, entropy models a cognitive state of the language user (the readers and their media in a narrative, i.e. the narrator and the mediating figure) associated with the amount of uncertainty about the outcome of a linguistic and narrative event relative to the general model of the world; the surprisal value is context-dependent and associated with individual elements of the sequence unfolding online: “Surprisal can thus be thought of as reflecting state-by-state expectation, with inputs that move the model to unexpected points in space yielding high surprisal. Entropy, in turn, quantifies how likely each fully specified state of affairs that makes up the space of meaning is, given the current point in space. Entropy reduction is thus effectively a metric of end-state confirmation, where greater reduction in uncertainty about the propositions being communicated as true, i.e. greater confirmation of the communicated state of affairs, leads to greater surprisal, which approximates the distance of that transition, while entropy reduction reflects a change in the inherent nature of those states: the degree of certainty about the communicated state of affairs” (Venhuizen 2019a).

The confusion of entropy and surprise leads to the paralogisms of making strange/ostranienie/Verfremdung:

While surprisal and making-difficult are proportional, entropy accounts for a very different situation, because both reducing entropy and increasing entropy involve difficulty, effort (relative to the absolute value of the change).

Both gaining and losing confidence in the world - feeling that you understand how it works and then feeling that you do not - is laborious. (“The comprehension-centric perspective on entropy predicts that both a reduction and an increase in entropy result in an increase in processing effort”, Venhuizen et al. 2019: 9.) Both premise and minus premise involve the expenditure of energy. The greatest difficulty arises when the subject oscillates between certainty and uncertainty, as in the case of being provoked and provoking.

### 4.3. Problems with Corpus Compilation

The manual search conducted by the research assistant in Work Package 5 Annika Blietz rendered following results:

**Deutsches Textarchiv**:

https://www.deutschestextarchiv.de/
→ no novels with „terrorism“ in the title found → mostly historical works

**Scottish Corpus of Texts and Speech:**

https://www.scottishcorpus.ac.uk/search/?search=Search&word=terrorism&author=&gender=-+All&region=-+All&spoken=y&written=y&title=&yearfrom=&yearto=&search=Search

only reports from parliament

**The Corpus of modern Scottish writing:**

https://www.scottishcorpus.ac.uk/cmsw/search/?search=Search&word=terror&author=&gender=-+All&title=&genre%5B%5D=Administrative+prose&genre%5B%5D=Expository+prose&genre%5B%5D=Personal+writing&genre%5B%5D=Instructional+prose&genre%5B%5D=Religious+prose&genre%5B%5D=Verse%2Fdrama&genre%5B%5D=Imaginative+prose&genre%5B%5D=Journalism&yeargroup%5B%5D=1700-1750&yeargroup%5B%5D=1750-1800&yeargroup%5B%5D=1800-1850&yeargroup%5B%5D=1850-1900&yeargroup%5B%5D=1900-1950&yearfrom=&yearto=&search=Search

- results only for keyword “terror”

- results are not about terrorism but about earlier meanings of terror (fear, dominion and violence)

**Austrian Books online:**

https://search.onb.ac.at/primo-explore/search?institution=43ACC_ONB&vid=ONB&tab=default_tab&search_scope=ONB_gesamtbestand&mode=basic&displayMode=full&bulkSize=10&highlight=true&dum=true&query=any,contains,terrorism&displayField=all

- many results for “terrorism” and more filter options for “International terrorism”, “USA”, “Fighting terrorism” and more

**Project Gutenberg:**

https://www.gutenberg.org/ebooks/results/

subject = terrorism -> 10 results (novels)

**Project Gutenberg AUS:**

https://gutenberg.net.au/searchresults.html

When searching the website for terrorism, you get 188 results. However, the results are mostly texts where the phrase terrorism/terrorist occur, but not necessarily texts about terrorism

**Project Gutenberg DE:**

https://www.projekt-gutenberg.org/info/search/search.php

- the search for terrorism does not give results that seem to match
- since the entries are sorted by topic (e.g. politics and social, history etc.) one can certainly find some entries about terrorism, but not easily by search bar

**Txtlab Multilingual Novels:**

https://figshare.com/search?q=terrorism

Many entries although not many novels, rather journals

**Russkaia virtual'naia Biblioteka:**

https://rvb.ru/

ca. 20 results for query "terrorism (терроризм)"

**Hathi Trust Digital Library:**

https://babel.hathitrust.org/cgi/ls?q1=terrorism&field1=ocr&a=srchls&ft=ft&lmt=ft

many results, ca 72.000 book results. Many filter options to filter the topic even more.

(according to open texts: 1429 results)

**Biblioteca Virtual Miquel de Cervantes:**

https://www.cervantesvirtual.com/buscador/?q=terrorismo&tipo=palabras_todas&busqueda=combinada&ftipo=texto

1050 results for "terrorism" with filter option "text"

**Europeana:**

https://www.europeana.eu/de/search?page=1&qf=TYPE%3A%22TEXT%22&query=Terrorismus&view=list

1238 results for terrorism, but mostly essays and non-fiction works

**Polona:**

https://polona.pl/sets?availableOnline=true&searchCategory=objectSets&searchLike=terroryzm

62 results for "terroryzm", but no filter for novel, mostly articles and non-fiction works

**Gallica:**

https://gallica.bnf.fr/services/engine/search/sru?operation=searchRetrieve&version=1.2&startRecord=0&maximumRecords=15&page=1&query=%28gallica%20all%20%22terrorisme%22%29&filter=dc.type%20all%20%22monographie%22%20and%20sdewey%20all%20%2280%22

959 results for "terrorisme" and the filter "literature"

**DIKDA:**

https://dikda.snk.sk/search?q=terorizmus%20&licences=paying_users,,public&doctypes=monograph

48 results for "terorizmus" and the filter "book"

**Univerzitná Knižnica v Bratislave:**

http://digitalna.kniznica.info/browse

2 results for search keyword "terorizmus" and filter option "book"

**Library of Congress:**

https://www.loc.gov/collections/world-digital-library/?q=terrorism

4 results (books/printed material)

**Federacja Bibliotek Cyfrowych:**

https://fbc.pionier.net.pl/search#fq={!tag=dcterms_accessRights}dcterms_accessRights%3A%22Dost%C4%99p%20otwarty%22&fq={!tag=tech_type}tech_type%3Aksi%C4%85%C5%BCka&q=dc_subject%3A(terrorism)

27 book results

**Open texts:**

Internet archive books (library)

https://opentexts.world/search?q=terrorism&organisation=Internet%20Archive%20Books

187 results

**Digital NZ (library):**

https://opentexts.world/search?q=terrorism&organisation=DigitalNZ

22 results

**National library of Scotland (library):**

https://opentexts.world/search?q=terrorism&organisation=National%20Library%20of%20Scotland

2 results

Conclusion:

Finding texts, especially novels, about terrorism was not easy in most cases. Often the authors did not provide metadata about the topics, which made the search for topics difficult and time-consuming. In addition, the word terror, which was often an alternative result to terrorism, is in most cases not the same as terrorism, but an old word for fear and violence. In most corpora there are no precise filtering options, and corpora such as the Zlatý Fond SME corpus are only sorted by author, which makes searching for a topic tedious. Furthermore, most of the results for terror are non-fiction. In conclusion, we can say that searching for texts on a specific topic in the corpora listed in the CLS Infra corpus table does not yield results that are easy to work with. Some corpora not only have a large number of entries, but also good options for searching for something specific. On the other hand, many corpora do not have these filter options.

### 4.4. Proposed Solutions – Corpus Exploration Platform and Tools

Again, we hope that the creation of the Corpus Exploration Platform (CEP) will help with the task of building the corpus. However, now that our corpus is being created with a specific research question in mind, we see the real value of the toolbox, which should definitely include affordances for measuring surprise. Moreover, the need to search text collections not only for formal features

(as stated in the metadata) but also for topics becomes even more pronounced than in the case of haiku. The methods of CLS lead us to acknowledge the old truth of literary studies that forms and themes are abstractions of an indivisible whole, even when the actual object of study is a corpus of texts rather than a single masterpiece.

In order to represent knowledge as organized yet multidimensional information, CEP uses an ontology as a framework. A graph database that contains RDF-encoded representations of corpora and associated documents as Linked Open Data is used to maintain the ontology. The database also comes with a customizable set of patterns for connecting such data to outside reference sources. Using ResearchSpace's front-end development template system, a dynamic knowledge graph of this nature can be expressed in a variety of data interactions, visualisations, rationalisations, and workflows. The ontology for CEP is designed as a descriptive metamodel for literary corpora, offering a controlled vocabulary for classifying the corpora, documents, and files, enabling a fine-grained description of collections according to a specific genre, including a list of formats, methods, tools, features, and topics. By recording provenance information, CEP will allow even conflicting statements about a resource to be made, which can be traced and evaluated individually. The latter is particularly important in the case of topics whose identification appears to be observer-related (subjective). It is therefore important to triangulate the CEP with wikidata and other RDF resources that are likely to contain information on the subject of a work.

At this stage, it is also important to work closely with WP8 to ensure that such mathematically advanced tools as those measuring surprisal and KD-divergence are accessible to literary scholars.

## 5. Conclusion and outlook: Corpus Exploration Platform and Programmable Corpora

In order of the complexity of the study subject, this deliverable presented three case studies involving digitalization and transformation processes. The case studies in this deliverable included: 1. Building an ELTeC affine corpus of the Slovak novel; 2. discovering the haiku in multilingual corpora; and 3. Measuring entropy and surprise in Tsarist Empire (Russian and Polish) prose devoted to terrorism.

The Slovak novel is a regional genre connected to a somewhat undeveloped Central European culture. Due to resource constraints, the first use case focuses on issues with resource accessibility and discoverability. In the long run, this use case also calls into question whether NLP technologies are universal and whether they are accessible for smaller languages. In case two, the haiku serves as an illustration of a global form typical of high modernism in Western literary cultures, in contrast to the Slovak novel. The haiku case introduces us to the world of cultural incommensurability, which combines with the disorder of metadata and corpus architectures. The issue now is not a lack of resources but rather the accuracy of browsing tools, namely how to locate the exact information we want in vast and varied corpora. The third case study, which is focused on the terrorism novel in Polish and Russian literature (as variations on the literature written under the conditions of the Tsarist empire), connects issues of accessibility and findability in a multilingual environment of less widely spoken languages to more focused research questions that could actually advance narratology. We track the effects of a real research hypothesis on corpus elaboration.

The Corpus Exploration Platform (CEP), which offers a technical solution to the issues faced by researchers attempting to explore today's data landscape, is presented in the second case and third case (see sections 3.3 and 4.4), whereas the first case reveals primarily hidden political and economic underpinnings of the digital humanities field that could easily be addressed with additional resources. While the second case study emphasizes the most fundamental and

essential qualities of CEP, the third one suggests options for future improvement of both resource cataloging techniques and natural language processing technologies.

The use cases can also serve as a starting point for thinking about programmable corpora (see CLS Deliverable 7.1), which is the ultimate goal of the technical and theoretical side of the CLS INFRA project. Programmable corpora are defined as corpora that provide an open, transparently documented and (at least partially) research-driven API for making texts machine-actionable. Each use case addresses a particular operation endpoint, as defined by the Distributed Text Services: The Slovak Novel corresponds mostly to the collections endpoint, which supports navigation across texts, while the focus on formal features in use cases two and three correlates with the navigation endpoint (which supports navigation within a text) as well as the documents endpoint. The latter enables the retrieval of complete or partial texts, thus providing an ultimate experience of literary study where close reading meets statistical data.

**References:**

Bartoszyński, Kazimierz (1971): Zagadnienie komunikacji literackiej w utworach narracyjnych. In: J. Sławiński (ed.), *Problemy socjologii literatury*. Wrocław: Ossolineum.

Brouwer, Harm & Francesca Delogu & Noortje J. Venhuizen & Matthew W. Crocker (2021): Neurobehavioral Correlates of Surprisal in Language Comprehension: A Neurocomputational Model. In: *Frontiers Psychology. Section: Language Sciences* 12.

Caracciolo, Marco (2014): *The Experientiality of Narrative. An Enactivist Approach*. Berlin & al.: De Gruyter.

Caracciolo, Marco (2016): Cognitive Literary Studies and the Status of Interpretation: An Attempt at Conceptual Mapping. In: *New Literary History: A Journal of Theory and Interpretation* 47(1), 187-207.

Certeau, Michel de (1984): *The Practice of Everyday Life*, translated by Steven Rendall. Berkeley: University of California Press.

Clark, Andy (2015): Embodied Prediction. In: Thomas *Metzinger* & Jennifer M. *Windt (ed.), Open MIND 7*: Frankfurt: The MIND Group, doi: 10.15502/9783958570115

Clark, Andy (2017): *Surfing Uncertainty. Prediction, Action, and the Embodied Mind*. Oxford: Oxford University Press.

Debnár, Marek & Jesypenko, Dmytro (2020): Budovanie komplexných a reprezentatívnych digitálnych literárnych zbierok v rámci Európskej zbierky literárnych textov (ELTeC) na Slovensku a Ukrajine. In: *Slovenská literatúra* 6, 630-638.

Degaetano-Ortlieb, Stefania & Andrew Piper (2019): The Scientization of Literary Study. In: *Proceedings of the 3rd Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, Minneapolis 18–28.

Degaetano-Ortlieb, Stefania & Elke Teich (2022): Toward an optimal code for communication: The case of scientific English. In: *Corpus Linguistics and Linguistic Theory* 18(1), 175-207. https://doi.org/10.1515/cllt-2018-0088

Degaetano-Ortlieb, Stefania (2019): Hybridization effects in literary texts. In: *Proceedings of the 10th International Corpus Linguistics Conference*, Cardiff.

Friston, Karl & Klaas Stephan (2007): Free-energy and the brain. In: *Synthesis* 159(3), 417-458.

Haider, Thomas (2021): "Metrical Tagging in the Wild: Building and Annotating Poetry Corpora with Rhythmic Features"*. Proceedings of the European Association for Computational Linguistics*, arXiv:2102.08858

Helmholtz, Hermann (1867): *Handbuch der physiologischen Optik*. Leipzig: Leopold Voss.

Hohwy, Jakob (2013): *The Predictive Mind*. Oxford: Oxford University Press.
Hokenson, Jan Walsh (2007): Haiku as a Western Genre. Fellow-Traveller of Modernism. In: Astradur Eysteinsson & Vivien Liska (eds.), *Modernism*. Amsterdam/Philadelphia: John Benjamins, vol. 2, 693-714.
Irvine, Gregory (2014): *Der Japonismus und die Geburt der Moderne. Leipzig:* Seemann Henschel.
Irving, Dan (2016): Presence, Kinesic Description, and Literary Reading. In: *CounterText* 2(3), 322-337.
Jakobson, Roman ([1921] 1969): Über den Realismus in der Kunst. In: J. Striedter & W. Stempel (ed.), *Texte der russischen Formalisten*. München, Bd. 1, 372-391.
Johnson, Jeffrey (2011): *Haiku Poetics in Twentieth-Century Avant-Garde Poetry*. Langham et al.: Lexington Books.
Kukkonen, Karin & Marco Caracciolo, ed. (2014): Cognitive Literary Study: Second Generation Approaches [Special Issue]'. In: *Style* 48(2).
Kukkonen, Karin (2014): Presence and Prediction: The Embodied Reader's Cascades of Cognition. In: *Style* 48(3), 367-384.
Kukkonen, Karin (2016): Bayesian Bodies: The Predictive Dimension of Embodied Cognition and Culture. In: Peter Garratt (ed.) *The Cognitive Humanities: Embodied Mind in Literature and Culture*. Basingstoke: Palgrave Macmillan.
Kukkonen, Karin (2018): The Fully Extended Mind. In: Zara Dinnen & Robyn Warhol (ed.), *The Edinburgh Companion for Contemporary Narrative Theory*. Edinburgh: Edinburgh University Press, 56-66.
Kukkonen, Karin (2019): *4E Cognition and Eighteenth-Century Fiction: How the Novel Found its Feet*. Oxford: Oxford University Press.
Levy, Roger (2015): Memory and surprisal in human sentence comprehension, https://www.mit.edu/~rplevy/papers/levy-2013-memory-and-surprisal-corrected.pdf
Long, Hoyt & Richard J. So (2016): Literary Pattern Recognition: Modernism between Close Reading and Machine Learning. In: *Critical Inquiry* 42 (Winter 2016): 235–67.
Lotman, Jurii ([1971] 1977): *The Structure of the Artistic Text*, translated by Ronald Vroon. Ann Arbor: University of Michigan Press.
Lowder, Matthew & Wonii Choi & Fernanda Ferreira, John M. Henderson (2018): Lexical Predictability During Natural Reading: Effects of Surprisal and Entropy Reduction. In: *Cognitive Science* 42 (4): 1166-1183. doi: 10.1111/cogs.12597.
Luxemburg, Rosa (1897): Der Sozialimsus in Polen. In: *Socialistische Monatshefte* 10, 547–556.
Odebrecht, Carolin (2018): *MKM – ein Metamodell für Korpusmetadaten. Dokumentation und Wiederverwendung historischer Korpora*, Dissertation. Humboldt-Universität zu Berlin, Sprach- und literaturwissenschaftliche Fakultät, Berlin. doi:https://doi.org/10.18452/19407
Olson, Greta, ed. (2011): *Current Trends in Narratology*. Berlin: De Gruyter.
Paris, Rainer (2015): *Der Wille des Einen ist das Tun des Anderen. Aufsätze zur Machttheorie*. Weilerswist-Metternich: Velbrück Wissenschaft.
Parr, Thomas & Giovanni Pezzulo & Karl J. Friston (2022): *Active inference: the free energy principle in mind, brain, and behavior*. Cambridge; MIT Press.
Patyk, Lynn (2018): Dostoevsky's Terrorism Trilogy. In: Peter Herman (ed.), *Terrorism and Literature*. Cambridge: Cambridge University Press.
Patyk, Lynn (2023): *Dostoevsky's Provocateurs*. Chicago: Northwestern University Press.
Pichler, Axel & Nils Reiter (2022). From Concepts to Texts and Back: Operationalization as a Core Activity of Digital Humanities. In: *Journal of Cultural Analytics*, vol. 7, no. 4, doi:10.22148/001c.57195.

Popova, Yanna (2015): *Stories, meaning, and experience. Narrativity and enaction*. New York & London: Routledge.
Richter, Ludwig (1979): *Slowakische Literatur: Entwicklungstrends vom Vormärz bis zur Gegenwart*. Berlin: Akademie.
Sayeed, Asad & Stefan Fischer & Vera Demberg (2015): Vector-space calculation of semantic surprisal for predicting word pronunciation duration. In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, Beijing, vol. 1, 763–773.
Schmid, Alex P. (2012): Revised Academic Consensus Definition of Terrorism. In: *Perspectives on Terrorism* 6(2).
Schmid, Wolf (2017): *Mentale Ereignisse. Bewusstseinsveränderungen in europäischen Erzählwerken vom Mittelalter bis zur Moderne*. Berlin & New York: De Gruyter.
Seth, Anil (2021): *Being You. A New Science of Consciousness*. London: Faber & Faber.
Schöch, Christoph & Roxana Patras & Tomaž & Diana Santos (2021): Creating the European Literary Text Collection (ELTeC) In: *Challenges and Perspectives. Modern Languages Open*, 0(1), .DOI: https://doi.org/10.3828/mlo.v0i0.364
Shannon, Claude Elwood (1948): *A Mathematical Theory of Communication. In: Bell System Technical Journal 27 (3), 379–423.*
Sławiński, Janusz (1971): Socjologia literatury a poetyka historyczna. In: Janusz Sławiński (ed.), *Problemy socjologii literatury*. Wrocław: Ossolineum, 65-78.
Šmatlák, Stanislav & Vladimír Petrík & Ludwig Richter (2003): *Geschichte der slowakischen Literatur und ihrer Rezeption im deutschen Sprachraum*. Bratislava: Literárně informačně centrum.
Śniecikowska, Beata (2016): *Haiku po polsku. Genologia w perspektywie transkulturowej*. Toruń: Wydawnictwo UMK.
Śniecikowska, Beata (2021): *Transcultural Haiku*. Berlin, Germany: Peter Lang Verlag.
Solms, Mark (2021): *The Hidden Spring: A Journey to the Source of Consciousness*. New York: Profole Books.
Venhuizen, Noortje J & Matthew W. Crocker & Harm Brouwer (2019): Expectation-based comprehension: Modeling the interaction of world knowledge and linguistic experience. In: *Discourse Processes 56*(3), 229–255. https://doi.org/10.1080/0163853X.2018.1448677
Venhuizen, Noortje J & Matthew W. Crocker & Harm Brouwer (2019a): Semantic Entropy in Language Comprehension. In: *Entropy 21*, https://doi.org/10.3390/e21121159
Venhuizen, Noortje J. & Matthew W. Crocker & Harm Brouwer (2019): Expectation-based Comprehension: Modeling the Interaction of World Knowledge and Linguistic Experience. In: *Discourse Processes* 56:3, 229-255, DOI: 10.1080/0163853X.2018.1448677
Venhuizen, Noortje J. & Petra Hendriks & Matthew W. Crocker & Harm Brouwer (2022): Distributional formal semantics. In: *Information and Computation* 287.
Wilkinson, Mark D. & Michel Dumontier & IJsbrand J. Aalbersberg (2016): The FAIR Guiding Principles for scientific data management and stewardship. In: *Sci Data* 3, 160018. doi: https://doi.org/10.1038/sdata.2016.18
Winiarski, Léon (1967): *Essais sur la mécanique sociale*, ed. by Giovanni Busino. Genève: Librairie Droz.
*Yasuda, Kenneth (*[1957] 2001): *The Japanese haiku*. Boston: Tuttle.
Zajac, Peter (1996): *Auf den Taubenfüßchen der Literatur. Ein Buch über slowakische Literatur und Kultur.* Blieskastel: Gollenstein.