# CineLexi: A corpus-based bilingual resource of evaluative vocabulary in film reviews (German–Russian)

**Mariia Kopylova**
Roma Tre University
Rome, Italy
mashulya.kopylova@list.ru

## Abstract

Evaluative vocabulary constitutes the core of the film review genre; however, existing CALL resources primarily focus on general vocabulary, and the domain-specific evaluative vocabulary of specialized genres such as film reviews remains largely unaddressed. Two independent gaps motivate this study: (i) general sentiment lexicons cover only a small subset of the functional-evaluative vocabulary that learners need for productive writing tasks, and (ii) for German, open CEFR-tagged lists are available only up to level B1, whereas the evaluative vocabulary attested in film reviews is predominantly at B2–C1.

The article presents CineLexi[1], a pilot, proof-of-concept corpus-based resource of evaluative vocabulary derived from film reviews for the German–Russian language pair. The corpus comprises 900 user reviews of 15 films selected from the IMDb Top 250 (5 genres × 3 films × 30 reviews × 2 languages), collected from the Kinopoisk (kinopoisk.ru) and Filmstarts (filmstarts.de) platforms. An eight-stage automatic candidate extraction pipeline including a manually curated exclusion list facilitated the extraction of 38 lemmas (18 German, 20 Russian) at proficiency levels B1–C1, each accompanied by normalized frequency data, examples, collocations, and chunks. The resource contains monolingual dictionary entries in both languages and a cross-lingual comparison section organized by evaluative categories rather than by direct translation equivalents.

Cross-validation using SentiWS and RuSentiLex shows that common sentiment resources cover only 23% of the final 38 lemmas (6/18 German; 3/20 Russian). The study identifies corpus-observed differences between the German Filmstarts and Russian Kinopoisk subcorpora: the German subcorpus shows a stronger concentration of craft-related evaluative vocabulary (cf. Bordwell and Thompson, 2008), while the Russian subcorpus shows a stronger concentration of narrative and existential vocabulary (cf. Chatman, 1978). The final 38 lemmas form a curated core; the supplementary candidate list (85 lemmas: 34 German, 51 Russian) provides material for expansion (Section 7). A pedagogical scenario for B2-level writing tasks is proposed but has not yet been empirically evaluated.

## 1 Introduction

Evaluative vocabulary is central to the film review genre; it is through this vocabulary that reviewers convey their judgments regarding a film's artistic quality, narrative structure, and emotional impact. However, this vocabulary remains largely unexplored within the context of foreign language teaching. Existing corpus-based dictionaries for computer-assisted language learning (CALL) tend to emphasize general language. Kilgarriff et al. (2014) developed frequency lists with CEFR annotations for nine languages "in a corpus-driven, principled way, but with the overriding priority of creating lists which are as useful as possible for language learners." Nevertheless, their source materials consist of general web corpora, and domain-specific evaluative vocabulary from specialized genres is not represented in these lists. Similarly, François et al. (2016) create SVALex as "a lexical resource primarily aimed at learners and teachers of Swedish as a foreign and second language that describes the distribution of 15,681 words and expressions across the Common European Framework of Reference (CEFR)."

This work is licensed under a Creative Commons Attribution 4.0 International Licence. Licence details: http://creativecommons.org/licenses/by/4.0/.

[1]All resources – including review identifiers, candidate extraction pipeline, dictionaries in CSV and XML formats, RelaxNG schema, HTML prototype of a dictionary entry, supplementary candidate list (85 lemmas), and Jupyter notebook – are publicly available at https://github.com/mariiakopylova/CineLexi under a CC BY 4.0 license (CEFR labels derived from GermanFluent: CC BY-SA 4.0).



*Proceedings of the 15th Workshop on Natural Language Processing for Computer Assisted Language Learning (NLP4CALL 2026)*
pages 244–255, November 19–20, 2026, University of Tartu Library

However, SVALex is derived from a corpus of educational texts (textbooks and teaching materials) rather than from authentic genre-specific discourse; film-review texts are not part of its source material.

The use of film texts in language learning is well supported by research literature. Webb (2010) analysed the scripts of 143 films (over one million word occurrences) and showed that repeated viewing provides multiple encounters with low-frequency vocabulary. Specifically, as the number of films analysed increased, so did the number of words encountered ten or more times – i.e., the relevant point is not raw frequency but repeated exposure across films. Sánchez-Auñón et al. (2023), in a systematic review, highlight the pedagogical potential of film as an authentic context for language learning. However, most of these studies focus on subtitles or dialogues, while user reviews as a source of evaluative vocabulary for computer-assisted language learning (CALL) remain virtually unexplored.

An additional challenge arises with the German–Russian language pair: unlike English–other language pairs, it is virtually absent from existing computer-assisted language learning (CALL) resources. For German, official open lists with CEFR tagging are available only up to level B1 (Goethe-Institut, 2016), whereas the evaluative vocabulary found in film reviews, as this study demonstrates, tends to be at levels B2–C1. Cross-validation against general sentiment lexicons – SentiWS (Remus et al., 2010) and RuSentiLex (Loukachevitch and Levchik, 2016) – reveals that only 23% of the candidate lemmas extracted in this study (6/18 DE; 3/20 RU) appear in these general resources. The functional-evaluative vocabulary of film reviews – words that assess a film through its aspects rather than direct polarity – systematically falls outside general sentiment resources, highlighting the need for a domain-specific lexicon.

Evaluative vocabulary is defined as lexical items that recurrently participate in linguistic constructions used to assess a film, its narrative, technical execution, emotional effect, or thematic significance, including both overtly evaluative expressions and aspect-denoting nouns that acquire evaluative function in genre-specific contexts. This article introduces CineLexi, a corpus-based vocabulary resource developed from film reviews for the German–Russian language pair. The corpus consists of 900 user reviews of 15 films from the IMDb Top 250 list, balanced across five genres: drama, sci-fi, crime/thriller, action, and comedy/romance. These reviews were collected from the Kinopoisk (kinopoisk.ru) and Filmstarts (filmstarts.de) platforms, with 450 reviews in each language. Using an automated pipeline with multi-stage filtering, 38 lemmas (18 German, 20 Russian) at proficiency levels B1–C1 were extracted and presented as monolingual dictionary entries. Each entry includes normalized frequency data, corpus examples, collocations, and productive multi-word chunks. Additionally, the resource features a cross-lingual comparison section organized by evaluative categories rather than by direct translation equivalents.

The central research question of this study is: What lexical profile characterizes the film review genre in German and Russian, and what cross-lingual asymmetries exist within this profile?

The cross-lingual differences observed in the corpus are reported in Sections 4.3 and 5.2.

Beyond the resource itself, the contribution to NLP4CALL is methodological: the pipeline combines spaCy lemmatisation and POS filtering, TF-IDF-based detection of title artefacts, regex-based frequency normalisation, and automatic CEFR tagging against GermanFluent, Goethe, and Kelly lists. The same pipeline can be transferred to other genres and language pairs, and the resource is designed to support keystroke-logging studies of L2 writing, in which learners' consultations of CineLexi entries can be logged and analysed at the level of evaluative category.

The article is organised as follows: Section 2 reviews the related literature; Section 3 describes the corpus and the candidate extraction pipeline; Section 4 presents the resource and key findings; Section 5 discusses the limitations and implications for CALL; Section 6 proposes a pedagogical scenario for using the resource; and Section 7 concludes the article.

## 2 Related Work

Corpus-based lexical resources for Computer-Assisted Language Learning (CALL) are developing along two primary lines: frequency lists with CEFR tagging and domain-specific dictionaries of evaluative vocabulary. These two lines are reviewed in Sections 2.1 and 2.2 respectively; Sec-

tion 2.3 then turns to film texts in language learning.

### 2.1 CEFR-graded lexical resources

The first approach is exemplified by the KELLY project (Kilgarriff et al., 2014) and the CEFR-Lex family – SVALex for Swedish (François et al., 2016) and EFLLex for English (Dürlich and François, 2018). SVALex describes the distribution of 15,681 Swedish words and expressions across the six CEFR levels (François et al., 2016), while EFLLex covers 15,280 English words in pedagogical materials (Dürlich and François, 2018). All three resources rely on educational or general web corpora; however, domain-specific vocabulary from specialized genres is not represented in them. The reliability of automatic CEFR tagging remains an open question: Alfter (2024) demonstrates that generative models grade vocabulary "with moderate but stable success", although their reliability for domain-specific units remains untested. Earlier, Alfter et al. (2016) showed that CEFR labels derived from different source corpora (general web vs. learner essays) yield different distributions, which suggests that frequency-based proxies derived from general corpora may overestimate the level of domain-specific vocabulary.

### 2.2 Sentiment lexicons

The second approach involves the use of sentiment lexicons. SentiWS (Remus et al., 2010) contains German words "weighted within the interval of $[-1; 1]$" and covers "not only adjectives and adverbs explicitly expressing a sentiment, but also nouns and verbs implicitly containing one"; RuSentiLex (Loukachevitch and Levchik, 2016) is compiled from "opinionated words from domain-oriented Russian sentiment vocabularies" and news collections. Both resources focus on direct polarity; however, functional-evaluative vocabulary – which assesses an object through its various aspects – is systematically absent, as suggested by the cross-validation reported in Section 5.

### 2.3 Film texts in language learning

Film texts in language learning are primarily examined through subtitles and dialogues (Webb, 2010; Sánchez-Auñón et al., 2023). Singh (2024) emphasizes the importance of corpus analysis of film texts for enhancing students' critical language awareness. However, user reviews – as a genre and as a source of evaluative vocabulary for computer-assisted language learning (CALL) – are not considered in these studies.

Corpus-based lexicography for computer-assisted language learning (CALL) is grounded in the principle of authenticity, as articulated by McEnery and Xiao (2011): authentic texts provide lexical contexts that are unavailable in traditional teaching materials. This study applies this principle to the genre of film reviews for the German–Russian language pair – a pair not represented in any of the referenced resources.

## 3 Corpus and Candidate Extraction Pipeline

### 3.1 Corpus

CineLexi is built on a corpus of user reviews in two languages. Russian reviews were collected from the Kinopoisk platform (kinopoisk.ru) via an unofficial API (poiskkino.dev), following the platform's terms of use for research purposes. German reviews were gathered from Filmstarts (filmstarts.de) using Selenium 4.52.0 (Selenium, 2026) with a headless Chrome browser, as the page content is dynamically loaded via JavaScript.

Fifteen films from the IMDb Top 250 list that appear on both platforms were selected, balanced across five genres (manually assigned based on platform metadata): drama ("The Shawshank Redemption," "The Green Mile," "Fight Club"), sci-fi ("Interstellar," "Inception," "The Matrix"), crime/thriller ("Pulp Fiction," "The Silence of the Lambs," "No Country for Old Men"), action ("Gladiator," "Léon," "The Dark Knight"), and comedy/romance ("Amélie," "The Grand Budapest Hotel," "Forrest Gump"). For each film, 30 reviews were collected in each language, totaling 900 texts (450 German and 450 Russian; 56,098 German tokens, 182,366 Russian tokens). Russian reviews are on average 3.4 times longer than German reviews (2,769 vs. 824 characters), necessitating frequency normalization per 1,000 tokens throughout the analysis. Table 1 summarizes the corpus statistics.

The corpus is a purposive sample: the 15 films were manually selected from the IMDb Top 250 and balanced across five genres (three films per genre). The repository does not redistribute the full review texts; it publishes review identifiers, extracted lexical data, and short illustrative frag-

| Metric | DE | RU |
|---|---|---|
| Reviews | 450 | 450 |
| Tokens | 56,098 | 182,366 |
| Avg. length (chars) | 824 | 2,769 |
| Films covered | 15 | 15 |
| Genres | 5 | 5 |
| Reviews per film | 30 | 30 |

Table 1: CineLexi corpus statistics.

ments. CEFR labels derived from GermanFluent are distributed under CC BY-SA 4.0; all other components are distributed under CC BY 4.0.

### 3.2 CEFR Reference Lists

CEFR annotation for German was conducted using three sources, prioritized as follows. The primary source is GermanFluent (germanfluent.com, CC BY-SA 4.0), a frequency-graded list of 14,931 lemmas covering all six CEFR levels from A1 to C2. This resource addresses the gap left by the official Goethe-Institut lists, which are publicly available only up to level B1. As a secondary cross-check, the Goethe-Institut Wortlisten (A1–B1) were used. For items absent from both sources, word frequency Zipf scores served as a fallback. For Russian, the Kelly word list (Kilgarriff et al., 2014; `ru_m3.xls` version) was the primary source, covering nouns and adjectives across CEFR levels A1–C2, with 5,925 entries retained after filtering to nouns and adjectives. Final CEFR coverage includes GermanFluent accounting for 88% of German entries in the core dictionary, while Kelly covers 100% of Russian entries.

### 3.3 Candidate Extraction Pipeline

The candidate extraction process involves eight successive stages, each guided by a specific methodological consideration. Table 2 shows the reduction of the candidate pool at each stage – separately for German and Russian. All thresholds listed below are heuristic and reflect the author's decisions; ablation studies are reserved for future work.

The first stage involves a part-of-speech (POS) filter. For lemmatizing all reviews, spaCy models of the medium (md) class (`de_core_news_md`, `ru_core_news_md`) were used. Subsequently, only nouns (NOUN) and adjectives (ADJ) were retained – two parts of speech that carry the primary evaluative content in film reviews. At this stage, stop words and tokens shorter than three characters were excluded. As a result, 3,895 German and 8,884 Russian lemmatized types were obtained.

| Stage | DE | RU |
|---|---|---|
| POS filter (NOUN + ADJ) | 3,895 | 8,884 |
| Frequency $\geq 5$ | 456 | 2,037 |
| Domain stopword removal | 430 | 2,012 |
| CEFR B1–C1 + anglicism filter | 233 | 1,201 |
| Film coverage $\geq 6/15$ | 107 | 724 |
| Filter A + Filter B + frequency norm + REMOVE | 18 | 34 |
| Top-20 selection – final dictionary | 18 | 20 |

Table 2: Candidate extraction funnel.

The second stage involves a frequency filter. Lemmas with a corpus frequency below five were discarded as statistically marginal; a word occurring fewer than five times in 900 reviews cannot be considered a stable candidate for a pedagogical dictionary. This threshold is consistent with the practice in corpus-based lexicography of discarding low-frequency items before applying association measures, since such measures are unstable on small counts (Church and Hanks, 1990; Evert, 2008). After applying this filter, 456 German and 2,037 Russian lemmas remained.

The third stage involves the removal of domain-specific stop words. A short list of generic nouns was manually compiled – words that are common in any film review but do not carry evaluative content (for example, *Film*, *Szene*, *Regisseur* in German; *фильм*, *история*, *режиссёр* in Russian). The list is deliberately minimal: only unambiguously generic domain words are excluded to avoid accidentally removing potentially evaluative vocabulary. After this stage, 430 German and 2,012 Russian candidates remained.

The fourth stage involves applying the CEFR filter (B1–C1) and removing Anglicisms. Candidates outside the B1–C1 range were excluded: items at levels A1/A2 are too simple for the target audience (students at level B1+), while items at level C2 are outside the target pedagogical band for this pilot resource. The B1–C1 range was selected as the target pedagogical band; CEFR levels are inherited labels, not independently validated proficiency levels. CEFR levels were assigned au-

tomatically based on reference lists, prioritizing GermanFluent → Goethe → wordfreq for German and Kelly → wordfreq for Russian. Additionally, Anglicisms were removed from the pool of German candidates. An Anglicism was identified using a frequency-based criterion: the difference between the English Zipf score and the German Zipf score of the same string exceeded 0.8. A stricter filter for transparent English cognates (Filter B) is described separately in Stage 7. This filter rests on a transfer assumption: a German lemma that is orthographically and semantically transparent through English is likely to be already known to learners with English proficiency, and is therefore not a productive learning target. This assumption is plausible for the German–Russian pair, where English is the dominant L2, but it has not been validated empirically; a learner study testing whether such items are actually known is part of future work. This stage yielded 233 German and 1,201 Russian candidates.

The fifth stage is the film coverage filter. The lemma must appear in reviews of at least 6 out of 15 films (40% of the corpus). Words concentrated in reviews of a single film tend to reflect thematic rather than genre-specific vocabulary – for example, *prison* appears almost exclusively in reviews of "The Shawshank Redemption." This filter reduced the pool to 107 German and 724 Russian candidates.

The sixth stage is Filter A, which detects title artifacts. TF-IDF was calculated based on film subcorpora, with each film treated as a single document. A word was identified as a title artifact if its maximum TF-IDF value across films exceeded the average by more than 8.0 times. This process removed 8 German and 80 Russian candidates – words whose frequency in the corpus is attributable to the subject matter of a single film rather than to genre-based evaluative usage.

Stage seven – filter B: detection of transparent internationalisms (German Only). This filter is applied exclusively to German candidates. A word was flagged if its umlaut-normalised form is attested in English with a Zipf frequency above 3.0; the intended orthographic-similarity check against a distinct English form was not effective in the current implementation (see Section 5.3). At this stage, 19 German candidates were removed. The complete list of identified internationalisms, along with their similarity coefficients and English Zipf frequencies, is available in the repository. It should be noted that Filter B focuses on English cognates and therefore does not cover Russian–German cognates. For example, the German word *genial* is transparent to a Russian-speaking learner through the Russian word *гениальный*, but it does not meet the criterion of English frequency. This limitation is discussed in Section 5.3.

The eighth stage involves frequency normalization and recalculation using regular expressions (regex). Since spaCy lemmatization and regex matching based on surface forms differ for inflectional variants, all frequencies were recalculated using language-specific regex patterns. For Russian, the patterns match word roots followed by the \w* extension (e.g., атмосфер\w* for *атмосфера*, \bфинал\w* for *финал*). For German, the patterns match the lemma in its capitalized form along with possible inflectional variations. A minimum normalized frequency threshold of 0.35 occurrences per 1,000 tokens was applied to filter out statistically marginal units, and a manually curated REMOVE list excluded regex false positives, overly general items, and intensifiers. Together with Filter A and Filter B, this reduced the candidate pool to 18 German and 34 Russian candidates.

The final stage is selection (top–20 per language). Of the 18 German and 34 Russian candidates remaining after Stage 8, the top 20 per language by regex frequency were retained (all 18 German items). The manually curated REMOVE list applied in Stage 8 comprised three categories: regex false positives, i.e., lemmas whose surface form matched a different, unrelated word (*быт* "daily life" matched forms of *быть* "to be"; *псих* "psych" matched *психологический* "psychological"; *благо* "good" matched *благодаря* "thanks to"; *поп* "priest" matched *попытка* "attempt"); items too general to be genre-specific, i.e., words that occur in any review but do not carry evaluative content (*момент* "moment", *ситуация* "situation", *уровень* "level"); and ordinal numerals misclassified as adjectives by spaCy (*первый* "first", *второй* "second", *третий* "third"). After selection, the final resource comprises 18 German and 20 Russian lemmas at CEFR levels B1–C1.

The candidate pool after Filters A and B, before the normalised-frequency threshold and the REMOVE list, comprising 34 German and 51 Russian lemmas (85 in total), is available at `https:`

//github.com/mariiakopylova/CineLexi.git. Units with known false positives from the regex require further validation before being used in pedagogical materials.

## 4 Description of the Resource

CineLexi is a corpus-based bilingual resource of evaluative vocabulary in the genre of film reviews for the German–Russian language pair, designed for use in computer-assisted language learning (CALL). The resource includes 38 dictionary entries – 18 in German and 20 in Russian – and a cross-lingual section organized not by direct translation equivalents but by evaluative categories. This structure reflects the differences in the evaluative strategies of the two languages and enhances their pedagogical accessibility.

### 4.1 Structure of the Dictionary Entry

Each CineLexi dictionary entry follows a standardized format and includes eight fields. The headword is presented as a lemma in its base form. The part of speech is determined automatically using spaCy; the final dictionary contains only nouns and adjectives, as these are the primary carriers of evaluative meaning in the review genre. The CEFR level is assigned automatically, with an explicit indication of the source: GermanFluent (CC BY-SA 4.0) for 88% of German entries, and Kelly (Kilgarriff et al., 2014) for 100% of Russian entries. All CEFR labels are indicative and have not been validated using student data (see Section 5.3).

The normalized frequency is presented as the number of occurrences per 1,000 tokens, separately for the German and Russian subcorpora. This normalization is necessary because the average length of a Russian review in the corpus is 3.4 times greater than that of a German review. The film coverage indicates how many of the 15 films in the corpus contain the given lemma; a minimum threshold of at least six films ensures that the word belongs to genre-specific vocabulary rather than the thematic vocabulary of a single film. Each article is accompanied by two corpus examples from reviews of different films (preference was given to short sentences no longer than 250 characters), up to three collocations – the most frequent content words immediately preceding the target lemma – and up to five chunks: frequent two- and three-word combinations that include the target word.

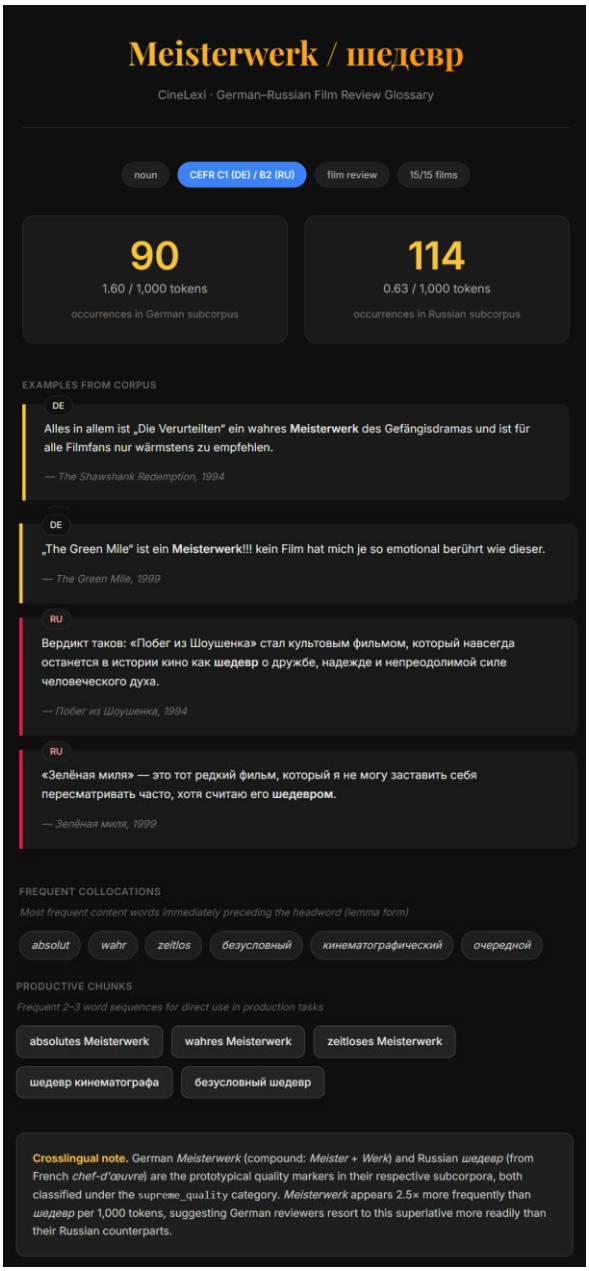



Figure 1: HTML renderer of the *Meisterwerk / masterpiece* article: eight CineLexi fields – headword, part of speech, CEFR level, frequency, film coverage, examples, collocations, chunks, plus a cross-lingual note.

Following Sinclair (1991), a collocation is treated as "the occurrence of two or more words within a short space of each other in a text", with the usual measure of proximity being "a maximum of four words intervening". A chunk, by contrast, is a frequent two- or three-word sequence containing the target lemma that functions as a ready-made production unit. The distinction is thus between a lexical association (collocation) and a prefabricated production unit (chunk). In the current implementation, collocations are oper-

ationalised more narrowly than in Sinclair's definition, as the most frequent content word immediately preceding the lemma; as a result, the two lists may overlap for adjective + noun patterns (e.g., *absolutes Meisterwerk*, *смысл жизни*), and a stricter separation is left for future work. Collocations and chunks are automatically extracted candidates and have not been manually validated; they are included in the resource as candidates rather than as validated pedagogical units. Known error types (over-generalising regex patterns, noisy preceding-word heuristics for adjectives) are documented in Section 5.3. Chunks are illustrated by such examples as *absolutes Meisterwerk* "absolute masterpiece", *düstere Atmosphäre* "gloomy atmosphere", *смысл жизни* "the meaning of life", and *человеческая природа* "human nature", intended for direct use in production tasks.

Figure 1 shows the HTML renderer of the dictionary entry *Meisterwerk* / *masterpiece* – an example of how eight fields are presented to the learner.

### 4.2 Final List of Lemmas

The complete list of 38 lemmas – with indications of the CEFR level, source of the label, corpus frequency, normalized frequency, and film coverage – is presented in the repository. In the main text, only a qualitative description of the distribution is provided. In the German sub-dictionary, lemmas at levels B2 (9 entries) and C1 (7 entries) predominate, which corresponds to the observation that evaluative lexis in German film reviews is predominantly at advanced levels, not covered by the official open lists of the Goethe-Institut. In the Russian sub-dictionary, levels B2 (11 entries) and B1 (7 entries) dominate – the distribution is lower, reflecting both the characteristics of the Russian subcorpus (longer analytical reviews with commonly used abstract vocabulary) and those of the Kelly list, which also covers more basic lexical units.

The most frequent lemmas in each language are indicative of the genre profile. In the German sub-dictionary, the core is formed by *Meisterwerk* "masterpiece" (1.60 per 1,000 tokens), *Darsteller* "actor" (0.94), *Spannung* "tension" (0.94), *Handlung* "plot" (0.89), and *Drehbuch* "screenplay" (0.66). The most frequent German items overall are the adjectives *perfekt* "perfect" (1.85) and *genial* "brilliant" (1.62). *Regie* "direction" (0.36) is also present in the final list as a craft-related item. The high-frequency discourse marker *Fazit* "conclusion" (1.19) is discussed separately in Section 4.3 as a genre-structuring item rather than a craft-evaluative noun. In the Russian sub-dictionary, the core is different: *сюжет* "plot" (2.65/1000t), *смысл* "meaning" (1.29), *драма* "drama" (0.97), *атмосфера* "atmosphere" (0.91), *идея* "idea" (0.89), *кадр* "shot/frame" (0.87), and *сценарий* "screenplay" (0.78) – predominantly nouns that evaluate narrative structure and thematic depth.

### 4.3 Cross-Lingual Mapping

A key feature of CineLexi is the <crosslingual> block – a structure that organizes lemmas in both languages not based on direct translation but according to genre assessment dimensions. Six such dimensions were identified, each described below. A complete representation of the categories, including lemmas and analytical notes, is provided in the resource's XML file.

Before the six dimensions are described, the two key analytical categories are defined. Craft is operationalised as the production-side dimensions of acting, directing, and screenwriting, following Bordwell and Thompson's (2008) account of film production. Narrative and existential vocabulary, by contrast, targets the story-side dimensions: plot structure, narration, thematic meaning, and fate (cf. Chatman, 1978).

The first dimension represents the highest quality rating. Both languages feature a frequently used noun denoting the pinnacle of artistic achievement: the German term *Meisterwerk* "masterpiece" (a compound of *Meister* + *Werk*) and the Russian term *шедевр* "masterpiece" (borrowed from the French *chef-d'œuvre*) are the most common markers of the highest quality in their respective subcorpora.

The second dimension is narrative structure. The German subcorpus evaluates narrative using two distinct nouns: *Handlung* "plot" (plot as experienced) and *Drehbuch* "screenplay" (script as a crafted object). In contrast, Russian distributes this evaluation across three units: *сюжет* "plot" (encompassing both German senses of *Handlung* "plot as experienced" and *Drehbuch* "script as a crafted object"), *повествование* "narration" (focusing on the mode of narration), and *сценарий* "screenplay" (the text as an object of evaluation). This threefold Russian distinction has no direct

counterpart in German.

The third dimension – emotional and atmospheric impact. The German subcorpus exhibits a more nuanced emotional vocabulary, employing three nouns with distinct connotations: *Atmosphäre* "atmosphere" (atmosphere as mood), *Gefühl* "feeling" (feeling as a subjective experience), and *Spannung* "tension" (tension as a structural arc). In contrast, the Russian subcorpus primarily conveys this evaluative dimension through a single term, *атмосфера* "atmosphere", which covers the semantic range distributed across the three German nouns.

The fourth dimension – acting and craft – is explicitly evaluated by German reviewers, who use specialized terms such as *Darsteller* "actor", *Hauptdarsteller* "lead actor", and *Drehbuch* "screenplay" to assess acting and script quality. The Russian final list also contains craft-related items (*состав* "cast", predominantly in *актёрский состав* "cast"; *кадр* "shot/frame"; *сценарий* "screenplay"; *визуальный* "visual"), but no dedicated actor noun, partly because *актёр* "actor" was treated as a generic stopword.

The fifth dimension – meaning and thematic depth. Russian reviewers assess films using existential and philosophical categories such as *смысл* "meaning", *идея* "idea", and *судьба* "fate". Among the German final lemmas, only *Tiefe* "depth" is directly comparable; *Sinn* "sense/meaning", the closest counterpart of *смысл*, was removed by Filter B (Section 5.3).

The sixth dimension involves reviewing discourse markers. The German term *Fazit* "conclusion" and the Russian *итог* "conclusion" are functionally equivalent, as both serve to frame the reviewer's final evaluation. These units function as genre-structuring markers rather than content-evaluative elements; however, their high frequency underscores their pedagogical importance for learners producing film reviews.

### 4.4 Resource Storage Format

The resource is stored in three formats: CSV, XML, and HTML. The CSV file (`cinelexi_entries.csv`) contains a tabular representation of all 38 dictionary entries with the following fields: lemma, language, CEFR level, CEFR source, corpus frequency, normalized frequency, film coverage, two examples, up to three collocations, and up to five chunks. The XML file (`cinelexi_dictionary.xml`) has a two-block structure. The <entries> block contains monolingual entries, each with the fields <headword>, <pos>, <cefr>, <domain>, <frequency>, <examples>, <collocations>, and <chunks>. The <crosslingual> block contains six categories described above; each category includes links to the corresponding dictionary entries and an analytical note.

The XML structure is formally defined using the RelaxNG Compact schema (`cinelexi_schema.rnc`). All 38 entries and six cross-lingual categories are validated against this schema. The HTML renderer for a single dictionary entry (`cinelexi_entry_meisterwerk.html`) is published as a prototype of the end-user interface, as illustrated in Figure 1.

This structure reflects a fundamental choice to create a cross-lingually annotated monolingual resource rather than a translation dictionary. CineLexi does not aim to provide direct translation equivalents for evaluative vocabulary; instead, it highlights the differences in the evaluative strategies of the two languages and offers students a tool to develop interlingual pragmatic awareness.

## 5 Discussion and Limitations

### 5.1 Cross-Validation Using Sentiment Lexicons

Cross-validation against SentiWS and RuSentiLex is consistent with the expectation that aspect-denoting evaluative vocabulary is underrepresented in general sentiment lexicons. However, no baseline is available for comparison: it is not known what coverage random nouns and adjectives of comparable frequency would receive. The observed overlap should therefore be interpreted as suggestive rather than conclusive. Of the 38 final lemmas, only 9 (23.7%) occur in SentiWS or RuSentiLex; a lemma is considered covered if it appears in at least one of the two resources — 6 out of 18 German lemmas (33%) and 3 out of 20 Russian lemmas (15%). Moreover, two of the three RuSentiLex matches are tagged as neutral (*реальность*, *человеческий*), which is methodologically uninformative for the evaluation task. The only genuinely useful match is *шедевр*, which is tagged as positive.

The low overlap suggests that general sentiment lexicons capture a substantially different subset of evaluative vocabulary from the one targeted by

CineLexi. The absence of items such as *Handlung*, *Drehbuch*, *сюжет*, *сценарий*, *итог*, and *Fazit* from both resources can be attributed to the nature of these words: they belong to functional-evaluative vocabulary, which assesses an object through its various aspects rather than through direct polarity. This particular class of words forms the core of CineLexi and is systematically absent from resources derived from news and web corpora. The observed coverage of 23% is therefore best interpreted as an indication that general sentiment lexicons and domain-specific evaluative resources capture different types of evaluativity, rather than as a measure of resource quality.

### 5.2 Asymmetry in Evaluative Strategies

The distribution of lemmas across evaluative categories differs between the two subcorpora. In this corpus, the German Filmstarts subcorpus shows a stronger concentration of craft-related vocabulary (*Darsteller*, *Drehbuch*, *Regie*, *Spannung*), whereas the Russian Kinopoisk subcorpus shows a stronger concentration of narrative and existential vocabulary (*смысл*, *судьба*, *идея*, *повествование*). This difference should be interpreted as a distributional tendency within this specific corpus rather than as a categorical opposition: several individual lemmas (e.g., *сценарий*, *визуальный*, *кадр*) pattern differently from the aggregate trend. This pattern may reflect differences in linguistic and/or platform-specific reviewing conventions. Kinopoisk encourages extended analytical reviews (average length: 2,769 characters), whereas Filmstarts favors shorter ones (824 characters).

The pedagogical significance of this asymmetry is considerable: a hypothesis is that a student who applies Russian-language assessment patterns to German may produce pragmatically marked texts, even at an advanced level of language proficiency. CineLexi makes this asymmetry explicit and pedagogically accessible through the <crosslingual> block, as well as through the instructional scenario proposed in Section 6.

### 5.3 Limitations

The first limitation is the size of the corpus. While 900 reviews for 15 films provide a sufficient volume for a pilot study, this sample is not large enough to support generalizations at the genre level. Expanding the corpus to include additional films, genres, and time periods is a priority for future iterations of the resource. Additionally, the 15 films represent a purposive sample, manually selected from the IMDb Top 250. This selection limits the external validity of conclusions regarding genre asymmetry, which should be interpreted as an observed trend within this specific corpus rather than a universal statement about the two languages.

The second limitation concerns the preliminary status of the CEFR labels. All CEFR levels in CineLexi are assigned automatically based on frequency lists and have not been validated using student data. This approach aligns with the practice of the CEFRLex family (François et al., 2016), but it necessitates a caveat: the levels should be considered indicative rather than normative. As Alfter (2024) demonstrates, automatic level assignment grades vocabulary "with moderate but stable success," even when employing generative models; however, reliability is lower for domain-specific vocabulary. Therefore, the CEFR labels in CineLexi should be regarded as a starting point for developing learning materials rather than as validated pedagogical recommendations.

The third limitation pertains to regex patterns and false positives. Several candidates in the supplementary list (85 lemmas) contain regex false positives, i.e., cases where a lemma's surface form matched an unrelated word: *быт* "daily life" matched forms of *быть* "to be"; *псих* "psych" matched *психологический* "psychological"; *благо* "good" matched *благодаря* "thanks to"; and *поп* "priest" matched *попытка* "attempt". In each case, the regex pattern for the intended lemma over-generated matches, inflating its frequency. These units are included in the supplementary data and require manual verification before use.

The fourth limitation concerns Filter B. Filter B (transparent internationalisms) targets English cognates and therefore does not cover Russian–German cognates. For example, the German word *genial* "brilliant" is transparent to a Russian-speaking learner through the Russian word *гениальный* "brilliant", yet it does not meet the English frequency criterion. In addition, the current implementation of Filter B has a methodological limitation: the orthographic similarity between a German lemma and its English counterpart was approximated by comparing the lemma with its own umlaut-normalised form, which yields similarity scores close to 1.0 for most items. As a

result, Filter B effectively reduced to a frequency check against English, and some items were removed that are not transparent English cognates – for example, *Sinn* "sense/meaning" (attested in 10 films), a direct counterpart of Russian *смысл* "meaning". This means that the German half of the meaning-depth category may be underrepresented in the final resource. A corrected implementation comparing German lemmas against English forms, and a symmetric filter using Russian as the bridge language, remain tasks for future work.

The fifth limitation is platform-specific. Reviews on Kinopoisk and Filmstarts reflect the conventions of their respective platforms and audiences, rather than the German and Russian languages per se. The observed asymmetries may be partly attributed to platform-related factors (such as review length and user demographics) rather than solely to linguistic differences. Testing this hypothesis requires a parallel corpus from other platforms – for example, IMDb for both languages.

The sixth limitation concerns the automatically generated lexicographic content. Several collocations and chunks were extracted using a heuristic that takes the word immediately preceding the target lemma; for adjectives, this heuristic produces noisy results (e.g., *Sache* "thing" for *hervorragend* "outstanding", *gesund* "healthy" for *Spannung* "tension"). The prefix-based regex patterns also over-generalise in some cases: for example, the pattern for *роман* "novel" matches *романтики* "romance", and the pattern for *состав* "cast" matches *составить конкуренцию* "compete", inflating the counts for these lemmas. The corpus examples were selected from the first two films in which the lemma appeared, which in practice means that most examples come from "The Shawshank Redemption" and "The Green Mile." The POS field in the XML is currently set to "noun" for all entries, including adjectives. These issues affect the presentation of individual entries rather than the aggregate lexical profile and will be addressed in a future iteration of the resource.

## 6 Pedagogical Scenario

Here is an example of using CineLexi in the context of Computer-Assisted Language Learning (CALL) within the framework of the PPP (Presentation–Practice–Production) model, one of the fundamental models for organizing educational materials in foreign language teaching.

Audience: B2-level students studying German or Russian as foreign languages; the task is to write a review of a film they have watched in the foreign language.

Stage 1: Presentation. The teacher presents the students with a cross-lingual CineLexi unit (Figure 1 shows a rendering of one article). The students analyse which evaluative categories are present in both languages and which appear in only one. Key questions include: "How does the German reviewer evaluate the actors' performances? What word does he use for this?" (*Darsteller*, *Hauptdarsteller*). "How does the Russian reviewer discuss the meaning of the film?" (*смысл*, *идея*).

Stage 2: Practice. Students receive flashcards containing case examples from CineLexi with missing target lemmas and reconstruct them using collocations and chunks. For example, for *Handlung*, the context "baut die ___ Stück für Stück geschickt auf" and the collocations (*emotionale*, *verständliche*) provide sufficient clues for recovery. For *сюжет*, the context "___ фильм основан на" and chunks (*сюжетная линия*, *сюжетные повороты*) guide the reconstruction.

Stage 3: Production. Students write a film review in the language they are learning, incorporating at least three lemmas from CineLexi within the appropriate language assessment categories. For native Russian speakers learning German, it is mandatory to include at least one item from the "Craft" category (*Darsteller* / *Drehbuch* / *Regie*), which is less frequent in the Russian subcorpus.

Beyond the product-oriented PPP scenario described above, CineLexi can also support process-oriented research on L2 writing. In a keystroke-logging setup, learners write a film review in a controlled environment (e.g., using Inputlog or a browser-based logger) while CineLexi entries are available in a side panel. The logger captures the entire writing process, including pauses, revisions, deletions, and insertions. This enables investigation into when and how learners consult CineLexi: whether they retrieve a lemma during the planning stage, formulation, or only during revision; whether they copy a chunk verbatim or adapt it; and whether consultations correlate with longer or shorter pauses. Because CineLexi entries are tagged by evaluative category

(craft, narrative, meaning, atmosphere, discourse markers), the log can be annotated at the category level, providing a detailed picture of which evaluative dimensions learners struggle with during production. For the German–Russian language pair, this is particularly informative: a Russian-speaking learner writing in German may consult craft-related entries (*Darsteller*, *Drehbuch*, *Regie*) more frequently than a German-speaking learner writing in Russian, or may avoid them altogether, reproducing the asymmetry observed in the corpus. Such data would allow to test whether the cross-lingual asymmetry documented in this study also manifests in learner production, and whether targeted CineLexi prompts can mitigate it.

This scenario illustrates the unique pedagogical application of CineLexi's cross-lingual structure: the resource serves not only as a dictionary but also as a tool for developing interlingual pragmatic awareness.

From an NLP4CALL perspective, the scenario illustrates how a domain-specific lexical resource can be integrated into process-oriented writing research: the spaCy-based pipeline, TF-IDF filtering, regex normalisation, and CEFR tagging that produced the resource also make it possible to align logged consultations with evaluative categories, and thus to connect NLP-derived lexical data with real-time learner behaviour.

The hypothesis to be tested is as follows: students who receive collocation exercises based on CineLexi will use more domain-appropriate evaluative vocabulary in their reviews than those in the control group. This scenario represents a proposal rather than an evaluated intervention; a pilot study involving B2-level students is planned as the next step.

## 7 Conclusion

This article presents CineLexi, a bilingual corpus-based resource of evaluative vocabulary in the film review genre for the German–Russian language pair. The corpus comprises 900 user reviews of 15 films (5 genres × 3 films × 30 reviews × 2 languages), collected from the Kinopoisk and Filmstarts platforms. An automatic pipeline consisting of eight automatic stages including a manually curated exclusion list enabled the extraction of 38 final lemmas (18 German and 20 Russian) at levels B1–C1. Each lemma is accompanied by normalized frequency data, corpus examples, collocations, and productive chunks.

Cross-validation with SentiWS and RuSentiLex suggests that general sentiment resources cover only 23% of the final candidates, motivating the need for specialized, domain-specific resources. In this corpus, reviews in the German Filmstarts subcorpus show a stronger concentration of craft-related vocabulary (*Darsteller*, *Drehbuch*, *Regie*), whereas reviews in the Russian Kinopoisk subcorpus show a stronger concentration of narrative and existential vocabulary (*смысл*, *судьба*, *повествование* – meaning, fate, narrative). This asymmetry has direct pedagogical implications for students learning German or Russian as foreign languages.

The study's limitations include the small corpus size (900 reviews for 15 films), the preliminary status of the CEFR labels, and the lack of validation using student data. Future plans include:

- expanding the corpus to cover more than 150 films, encompassing a wider range of genres and time periods
- manual verification of an additional list of candidate lemmas (85 in total) with the involvement of native speakers
- validation of the CEFR labels on student data
- conducting a pilot study of the pedagogical scenario described in Section 6.

The extraction pipeline, review identifiers, final dictionary (CSV, XML), RelaxNG schema, HTML prototype, and the supplementary candidate list are available in the repository. CEFR labels derived from GermanFluent are released under CC BY-SA 4.0; all other components under CC BY 4.0. The Russian CEFR reference list is based on Kelly (Kilgarriff et al., 2014).

## AI Disclosure

☐ Not applicable (no AI tools used)
☒ Assistive use
☒ Editing
☐ Generation of minor elements
☐ Translation
☒ Compression/expansion
☐ Generation from prompt

**Details:** AI tools were used for assistive purposes (idea generation, argument development), editing

(grammar, spelling, flow improvement), and compression/expansion (shortening or expanding existing human-generated input to meet page limits). All AI-assisted output was reviewed and edited by the author, who takes full responsibility for the final content.

## References


David Alfter. 2024. Out-of-the-box graded vocabulary lists with generative language models: Fact or fiction? In *Proceedings of the 13th Workshop on Natural Language Processing for Computer Assisted Language Learning (NLP4CALL 2024)*, pages 1–19, Rennes, France. LiU Electronic Press.

David Alfter, Yuri Bizzoni, Anders Agebjörn, Elena Volodina, and Ildikó Pilán. 2016. From distributions to labels: A lexical proficiency analysis using learner corpora. In *Proceedings of the Joint Workshop on NLP for Computer Assisted Language Learning and NLP for Language Acquisition at SLTC*, Umeå, Sweden. LiU Electronic Press.

David Bordwell and Kristin Thompson. 2008. *Film Art: An Introduction*, 8th edition. McGraw-Hill, New York.

Seymour Chatman. 1978. *Story and Discourse: Narrative Structure in Fiction and Film*. Cornell University Press, Ithaca.

Kenneth W. Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29.

Luise Dürlich and Thomas François. 2018. EFLLex: A graded lexical resource for learners of English as a foreign language. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Stefan Evert. 2008. Corpora and collocations. In A. Lüdeling and M. Kytö, editors, *Corpus Linguistics: An International Handbook*, volume 2, pages 1212–1248. Mouton de Gruyter, Berlin.

Filmstarts. 2026. Filmstarts user reviews. filmstarts.de. Accessed September 2026.

Thomas François, Elena Volodina, Ildikó Pilán, and Anaïs Tack. 2016. SVALex: A CEFR-graded lexical resource for Swedish foreign and second language learners. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 213–219, Portorož, Slovenia. European Language Resources Association (ELRA).

GermanFluent. 2026. GermanFluent: Frequency-graded German vocabulary lists (A1–C2). germanfluent.com. CC BY-SA 4.0. Accessed September 2026.

Goethe-Institut. 2016. *Goethe-Zertifikat B1: Wortliste*. Goethe-Institut, München.

ilkermeliksitki. 2023. Goethe Institute word list (GitHub mirror). github.com/ilkermeliksitki/goethe-institute-wordlist. Accessed September 2026.

Adam Kilgarriff, Frieda Charalabopoulou, Maria Gavrilidou, Janne Bondi Johannessen, Saussan Khalil, Sofie Johansson Kokkinakis, Robert Lew, Serge Sharoff, Ravikiran Vadlapudi, and Elena Volodina. 2014. Corpus-based vocabulary lists for language learners for nine languages. *Language Resources and Evaluation*, 48(1):121–163.

Kinopoisk. 2026. Kinopoisk user reviews. kinopoisk.ru. Accessed September 2026. Reviews collected via unofficial API: poiskkino.dev.

Natalia Loukachevitch and Anatolii Levchik. 2016. Creating a general Russian sentiment lexicon. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1171–1176, Portorož, Slovenia. European Language Resources Association (ELRA).

Tony McEnery and Richard Xiao. 2011. What corpora can offer in language teaching and learning. In E. Hinkel, editor, *Handbook of Research in Second Language Teaching and Learning*, volume 2, pages 364–380. Routledge, London and New York.

Robert Remus, Uwe Quasthoff, and Gerhard Heyer. 2010. SentiWS – a publicly available German-language resource for sentiment analysis. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, pages 1168–1171, Valletta, Malta. European Language Resources Association (ELRA).

Estefania Sánchez-Auñón, Pedro Antonio Férez-Mora, and Fuensanta Monroy-Hernández. 2023. The use of films in the teaching of English as a foreign language: A systematic literature review. *Asian-Pacific Journal of Second and Foreign Language Education*, 8(1):10.

Selenium. 2026. Selenium WebDriver (version 4.52.0). selenium.dev. Accessed September 2026.

John Sinclair. 1991. *Corpus, Concordance, Collocation*. Oxford University Press, Oxford.

Manvender Kaur Miriam Singh. 2024. Corpus-based cultivation of critical language awareness: A case study on English film subtitles. *Journal of Interdisciplinary Studies in Education*, 13(S1).

Stuart Webb. 2010. A corpus driven study of the potential for vocabulary learning through watching movies. *International Journal of Corpus Linguistics*, 15(4):497–519.