The role of subjectivity in lexicography: Experiments towards data-driven labeling of informality

Lydia Risberg[1, 2], Eleri Aedmaa[1], Maria Tuulik[1], Margit Langemets[1], Ene Vainik[1], Esta Prangel[1], Kristina Koppel[1], Hanna Pook[1]

[1] Institute of the Estonian Language, Roosikrantsi 6, Tallinn, Estonia
[2] University of Tartu, Jakobi 2, Tartu, Estonia
E-mail: lydia.risberg@eki.ee, eleri.aedmaa@eki.ee, maria.tuulik@eki.ee, margit.langemets@eki.ee, ene.vainik@eki.ee, esta.prangel@eki.ee, kristina.koppel@eki.ee, hanna.pook@eki.ee

Abstract

Language corpora have long been used in linguistics and lexicography, but recent developments now allow large language models (LLMs) to support or even transform these fields. This study investigates the potential of LLMs for annotating informal language use in Estonian – a language underrepresented in LLM training data yet supported by a large corpus. Focusing on the informal register label used in the *Dictionary of Standard Estonian*, we explore whether LLMs can assist lexicographers in determining the informal label. This paper describes two experiments that make use of LLMs, including GPT, Gemini, and Claude. The first experiment yielded useful insights but also highlighted necessary improvements. In the second experiment, we evaluated the LLMs' consistency and accuracy in categorizing words as informal or neutral/formal. Results showed that LLMs achieved around 76% agreement with expert human annotators, significantly above random chance, suggesting their usefulness as a supplementary resource in lexicography. GPT-4o demonstrated high accuracy, stability, and cost-efficiency, making it a reliable candidate for such a lexicographic task. The study highlights the inherent subjectivity in register labeling and the value of combining corpus data, expert judgment, and LLM output. Overall, LLMs represent a promising tool for modern dictionary work.

Keywords: large language models; register labels; Estonian; lexicography; informal language

1. Introduction

"Objects cannot always be assigned unambiguously to watertight categories, and instead we should think in terms of 'degrees of category membership'" (Rundell, 2002: 146). The same applies to words, or rather, their meanings: for many of them, it is not clear to what extent they are perceived as, for example, informal by different people – e.g., whether Estonian word *aastanumber* '(literally:) year number' is perceived as informal or not by majority of speakers (see Section 2.1 for details). This fuzziness is a

problem in lexicography, where clear decisions must be made about whether to assign a register label to a word meaning or not, and where one cannot afford to spend weeks or months analyzing a single word – the decision must be made as quickly and, ideally, as objectively as possible (at least in descriptive lexicography[1]).

Language corpora have long been used in linguistics and lexicography (Klosa-Kückelhaus & Tiberius, 2024; Baayen, 2024; Rundell, 2024), but now there is also the possibility of using large language models (LLMs). A wide range of experiments have been conducted with LLMs, many of them using the ChatGPT interface (e.g., de Schryver, 2023; Marcondes et al., 2024; Trap-Jensen, 2024; Rundell, 2024). The present study uses an API-based approach on different LLMs, and we focus on the Estonian language, which is underrepresented in LLMs training data (which shows a strong bias towards English; Li et al., 2024), but not a low-resourced language – the *Estonian National Corpus* (ENC, 2023) contains 3.8 billion tokens (Koppel et al., 2023).

Different labels are used in dictionaries to indicate that a word tends to be used outside a neutral context (e.g., *informal* or *colloquial*, *old-fashioned*, *poetic*, *offensive*).[2] These labels serve as a reference point for language users and language learners seeking register information (Langemets et al., 2024: 751). The words labeled in the normative *Dictionary of Standard Estonian* (DSE) have not previously been systematically studied. Moreover, some words shift over time from informal to neutral language, making it necessary to review the material before each new edition of the dictionary is published – the next DSE is scheduled for release in December 2025. Not all labels in the previous DSEs were based on research into actual language usage; some were instead based on decisions made by language planners. Therefore, preparing a new edition required reevaluating whether certain labels should be removed (or others added). The focus of our study is specifically on the label *informal* (or *colloquial*; in Est. *kõnekeelne*), which contrasts with neutral and formal language.

The assignment of labels has also, in the past, been influenced by the lexicographer's intuition (e.g., Karelson, 1990). However, register analysis should not be based on individual preferences (Biber & Conrad, 2009). Therefore, a systematic analysis of broader language use is useful when determining register labels. Although decisions in descriptive lexicography have been based on corpus analysis (see Section 2.1), in Estonia, there hasn't been a common guideline for lexicographers using the corpus query tool Sketch Engine (Kilgarriff et al., 2014) as to whether a word tends to be used in informal contexts.

[1] The Institute of the Estonian Language underwent a successful evaluation of research institutions, with particular praise given to our descriptive approach (see Regular Evaluation Report 2024: 1)

[2] Rundell (2024) makes the distinction between grammatical and sociolinguistic labels. A grammatical label indicates "that, for example, a particular verb has a strong preference for occurring in the passive or for not occurring in progressive forms. Sociolinguistic labels are applied to words or meanings whose distribution across text types is in some way limited."

In this paper, we introduce two experiments involving pre-trained LLMs. We examined whether LLMs can be used to analyze Estonian and make decisions regarding the informal label. We also discuss whether a lexicographer should ignore their linguistic intuition or whether, and to what extent, they should rely on it.

2. Background

First, a note on terminology. In theoretical discussions, authors have used terms such as *register* and *genre* sometimes to refer to the same phenomenon, and sometimes to distinguish between different ones. These concepts have inherently fuzzy boundaries and partially overlap. (Biber & Egbert, 2023; Biber & Conrad, 2009; Vaik, 2024) In this study, *register* is broadly understood as a scale of situational use, ranging from informal to neutral to formal. More specifically, language register is conceptualized as socially recurring intra-individual variation shaped by situational and functional contexts, with formality being a key dimension of these contexts (Lüdeling et al., 2022; Rotter & Liu, 2023).

Registers have come to be seen as continuous rather than clear-cut categories – that is, they can be recognized through certain linguistic markers or situational contexts, but they cannot be clearly delimited on this basis (Biber & Egbert, 2023; Henriksson et al., 2024: 308). From the perspective of the present study, the challenge lies in the fact that the structure of a dictionary isn't well suited to representing registers, since a word is presented apart from its natural context. Authentic example sentences illustrate word usage much more effectively than a label alone.

In our first experiment, the term *genre* was understood broadly as referring to a specific type of text, such as a blog post. Admittedly, this approach is overly general, as it reflects the classification of texts in the *Estonian National Corpus*. Texts serve different purposes and address audiences – some follow strict conventions, while others are more flexible. Accordingly, certain genres tend to lean toward informality (e.g., blog posts, forum comments), whereas others, such as legal texts and academic writing, lean toward formality and are typically subject to (strict) external linguistic control. However, since this paper does not engage with the corpus analysis dimension, we do not elaborate further on the conceptual scope of *genre*.

2.1 Estonian dictionaries and the informal label

Estonian lexicographers working on descriptive dictionaries have analyzed and evaluated the register of words based on language data and examples. For instance, the *Estonian National Corpus* (Koppel & Kallas, 2022) has been in use since 2010s, and the *EKI Combined Dictionary* (available since 2020 in the language portal *Sõnaveeb* 'Word web', see Koppel et al., 2019) has been compiled using corpus data. However, lexicographers of the Institute of the Estonian Language (EKI) have often found it

challenging to decide whether to add or remove register labels, as there has previously been no shared guideline on how to systematically use corpora for analyzing registers at the level of word meanings.

Furthermore, words with register labels in Estonian dictionaries have not yet been thoroughly and systematically reanalyzed. But labels in dictionaries are often interpreted as if they were objective information, not as subjective (Pajusalu, 2009: 6). In Estonian language sources – particularly older ones – labeling has often involved subjective decisions. As has been noted in relation to the compilation of the *Explanatory Dictionary of Literary Estonian*: "Despite everything, all of our labels indicating colloquial, humorous, derogatory, or other usage may turn out to be entirely subjective, since they are based mainly on the personal language intuition of the dictionary's authors" (Karelson, 1990: 33–34).

Secondly, the *Dictionary of Standard Estonian* is a normative dictionary. As such, some of its labels reflect the choices of language planners rather than descriptive observations: "The language form prescribed by the rules of usage guides and the DSE has not arisen naturally – it has been created. The language planners are people who make choices between competing variants and determine the norms." (Raag, 2008: 11–12; see Beal et al., 2023 for linguistic prescriptivism)

Thus, labeling practices have been influenced by subjective judgments and prescriptive choices. But over time, words can shift across register boundaries due to frequent and widespread use. For example, Estonian words *aku* 'battery' (from the longer *akumulaator*), *näts* 'chewing gum' (from *närimiskumm*) and *vildikas* 'felt-tip pen' (from *viltpliiats*) were labeled informal in the DSE 1976, but as their usage in neutral contexts became more common, they were no longer labeled informal in the 1999 edition. Nevertheless, not all words have been thoroughly reanalyzed, or language planning has not accepted the broader usage of certain words, so some labels have remained even though they no longer should have. Thus, the aim of the experiments described in this paper has been to contribute to the systematic reanalysis of words labeled as informal.

By analyzing why some words in the DSE carry an informal label (these words were selected from the second experiment; see Section 3.2), we found both reasons that seem unquestionable from a native speaker's perspective (examples a–c) and other cases that raise doubts (d–f).

a. the *-kas* suffix, as in the words *süümekas* 'guilty conscience' (from the longer compound word *süümepiin*) and *huulekas* 'lipstick' (from *huulepulk*)

b. shortened form of a word, such as *mill* 'million' (from *miljon*) and *vibra* 'vibration' (from *vibratsioon*)

c. words with a negative or pejorative connotation, such as the expletive *jeekim*

'goblin, little devil' or *porduelu* 'life of debauchery'

d. language planning traditions, including efforts to reduce linguistic variation (see Risberg 2024: pp. 100), e.g., neutral *igapidi* 'in every way' vs. "informal" *igatepidi*

e. figurative expressions, e.g., *sihverplaat* 'face' (literal meaning 'clock dial') and *kirvereegel* 'rule of thumb' (literal meaning would be 'rule of axe')

f. the word is just not a technical term, such as "informal" compound word *aastanumber* 'a number indicating a calendar year' vs. neutral term *aastaarv* – the distinction in specialized language is made between the words *number* 'symbol representing a number' (0–9) and *arv* 'mathematical value expressing quantity'. However, in general language, both *aastanumber* and *aastaarv* are used to denote a year number.

Interpretation difficulties have arisen because from the DSE 1999 to DSE 2018 the informal label was used in two meanings: written standard language versus informal language, and technical language versus general language (see critique by Vare, 2001; Kasik, 2021). The imposition of technical language needs onto general language in the previous DSEs is a topic that deserves its own treatment, but the relationship between terminology and general language is not the focus of this study as the next DSE in 2025 (and the *EKI Combined Dictionary*) are mainly general language dictionaries.

2.2 LLMs

LLMs are trained on massive amounts of data, the vast majority of which consists of texts collected from the web (Grattafiori et al., 2024; likewise, in the ENC, see Koppel & Kallas, 2022). The training data used in the register label experiments has not been curated with specific attention to the volume or quality of Estonian language content. Many (commercial) model developers do not disclose details about the training data or processes (Yu et al., 2024: 656).

Nevertheless, it is known that there are several limitations in applying these models. In the context of Estonian, one of the most significant issues – shared with many other languages – is that the Estonian training data is not comparable to the English data in terms of either volume or diversity (Nguyen et al., 2024: 3501). As a result, the language and content generated by the models may be influenced by the English language and cultural context.

Moreover, the training data contains stereotypes that can be reflected in the model's outputs (Bender et al., 2021; Tao et al., 2024), and since there is no full transparency regarding the data used, it is difficult to fully explain how the model arrives at specific results. The models may also generate information that is not factually accurate. This

reduces their reliability, and depending on the purpose, it may be necessary to validate the results. From a research perspective, model outputs may not be reproducible (Brown et al., 2020; de Schryver, 2023; Hadi et al., 2023). The externalities of LLM use include concerns about environmental impact and replication of bias (McKean & Fitzgerald, 2024).

3. Two experiments on pre-trained LLMs

We chose to analyze the informal label for practical lexicographical reasons. Capturing informal language has been more challenging for lexicographers because informal (i.e., colloquial) language as a register is more borderline and subjective than, for example, vulgar or child language. Yet the decision whether to include a label still has to be made. For example, compare the vulgar *pasarahe* 'shitstorm' or the child language *notsu* 'piggy' with complex cases like non-term *aastanumber* 'year number' and figurative *kirvereegel* 'rule of thumb', where the label is less justified. Selecting more extreme words would have been easier for evaluating LLMs, but we also wanted to gain practical benefit by reviewing words labeled as informal, since they needed to be systematically reviewed for DSE 2025 (see Section 2.1).

Next, we will describe the two experiments we conducted on pre-trained LLMs (in 2024 and 2025)[3] to support lexicographers' work with the informal label. We accessed LLMs via API, not via chat interfaces. Both experiments employed a zero-shot prompting (asking without examples) approach because our goal was to explore what the LLMs "know" without prior fine-tuning. We designed the tasks that simulate lexicographic work or assist lexicographers with informal labeling ourselves, because we didn't find previous studies that examined register labels as thoroughly with LLMs, especially in relation to Estonian. For example, Jakubíček and Rundell (2023: 527–528) did experiments with ChatGPT for lexicographic work, but labels were only a brief part of the task. There, ChatGPT identified one offensive and one archaic English word but failed to recognize a rare word (see Trap-Jensen, 2024 for critique).

3.1 First experiment in 2024

The first experiment revealed some shortcomings in our approach. Nonetheless, we will provide an overview of it here, as documenting and discussing areas for improvement adds value to scientific progress, especially in emerging fields like generative AI, where methods and best practices are still evolving (Tu et al., 2025).

Our first goal was to examine whether LLMs can be of any use for studying Estonian at all, since Estonian is underrepresented in LLMs' training data compared to English,

[3] All experimental materials are available at https://github.com/keeleinstituut/EKKD-III1/tree/main/registrid (16 September 2025).

although it is not a low-resourced language. At the time of the experiment, in March and May 2024, three model families demonstrated a reasonable ability to generate Estonian: OpenAI's GPT models (OpenAI, 2023), Google's Geminis (Gemini Team, Google, 2023), and Anthropic's Claude models (Anthropic, 2024). The Estonian proficiency of these and other LLMs had not been systematically evaluated, so this assessment was based on the subjective observations of our team members.[4] In general, newer models tend to perform better, but when selecting models, factors such as ease of implementation and cost should also be considered.

We prompted GPT-4 and Gemini 1.0 Ultra in March and GPT-4o and Claude Sonnet 3 in May 2024. We investigated whether these LLMs are capable of accurately determining the textual contexts in which Estonian words occur. As mentioned above, we chose the zero-shot approach. If we had directly asked whether a word is informal, we would have received a simple yes/no answer and would not have learned in which types of texts the LLM considers the word to be used. The role was described as "You are a compiler of an Estonian dictionary" and the prompt was: "In what kinds of texts is the given Estonian word used? If you don't have information about that, say that you don't know."[5] We selected material from two general language dictionaries: random 50 words labeled as informal from the normative DSE 2018[6] and 50 from the newer *EKI Combined Dictionary* 2024, plus an equal number of unlabeled words from each as a comparison. Additionally, we included 20 dialect words from each dictionary, as these too deviate from the neutral standard language.

Rather than presenting the results, we will discuss the issues with experiment design and outcomes. First, Claude and Gemini were excluded from the final experiment based on decisions made through superficial observation: e.g., Claude claimed that the Estonian word *mutt* 'mole' is a dog. However, it was not quantitatively assessed whether Claude and Gemini performed (statistically) significantly worse than the GPT models.

Also, we did not evaluate the outputs of GPT-4 and GPT-4o separately but rather grouped them together. So, when we assessed the informativeness of the LLM – that is, whether the information provided by the LLM enabled the lexicographer to determine which label to assign to the word – we found their combined informativeness to be 81%. However, we do not know what the result would have been if they had been evaluated separately as there were many comments left by lexicographers that GPT-4 performed worse, hallucinating more often than GPT-4o, which more frequently admitted when it didn't know a word. A (weak) correlation was found between the (higher) frequency of a word in the ENC 2023 and whether the LLM was able to say something meaningful about that word. This was particularly evident with dialect

[4] There are now ongoing efforts to assess the extent to which LLMs understand and generate Estonian language and culture, see https://baromeeter.tartunlp.ai/ (16 September 2025).

[5] Originally, the role and prompt were in Estonian; here is their English translation.

[6] In early 2024, the compilation of the DSE 2025 had not progressed as far as a year later.

words, given that dialect material is largely absent from the ENC and most likely also from LLMs.

Another issue with the design was that we were imitating the real work of a lexicographer who doesn't generally engage in lengthy discussions about a single word with various colleagues. This is why each word was assessed by only one lexicographer – having at least three experts assessing each word would have been better (see Section 3.2). A positive outcome of the experiment was that we discovered LLMs (especially GPT-4o) showed promise when applied to lexicographic tasks of this kind and were worth exploring further in experiments involving Estonian.

This experiment had a second part as well. Since we didn't know much about the capabilities of the LLMs in early 2024 and had no access to their training data, we conducted a corpus analysis for comparison, to see whether the data in the LLMs is reliable. As the design of the whole experiment needed improvements, we will not dwell on that analysis here. Nevertheless, we did focus more closely on the corpus data analysis (see Risberg et al., 2025). Here we will only note that the informativeness of ENC 2023 turned out to be 82.1%, which is roughly on par with that of the GPT models (81%).

3.2 Second experiment in 2025

Our team decided to continue with the same direction to contribute to the systematic reanalysis of words labeled informal for the upcoming DSE 2025. In the second experiment, the material consisted of words marked for inclusion in DSE 2025 that had at least one meaning labeled as informal. In late 2024, there were about 1,500 such words. This time, our goal was to explore practical ways of using LLMs to support lexicographers' decisions about the informal label.

3.2.1 Choosing LLMs and designing the prompt

The first step of this experiment was conducted in January 2025. We evaluated different LLMs available from OpenAI, Anthropic and Google at the time using a specially curated evaluation benchmark of 30 words and refined the prompt. Among these 30 words were both informal and neutral words (e.g., the informal shortened word *turva* versus the neutral *turvamees* 'male security guard'). The benchmark was established through discussion and analysis by all 9 team members regarding whether each word meaning is used mainly in informal contexts, and the LLMs' responses were compared against these judgments. As a result of this step, we selected the following models for the next step: 1) GPT-4o, 2) Gemini 1.5 Pro, 3) Claude Sonnet 3.5 and 4) Claude Sonnet 3.7 (which became available in February). We did not include the Meta's Llama models (Touvron et al., 2023) due to their continuous poor performance with Estonian at the time, nor DeepSeek (Bi et al., 2024) due to security concerns.

In February, the second step was to extract all DSE 2025 words from the database that

had at least one informal meaning. At the end of February, there were 1,330 such entries. We applied the role and task developed in January (originally in Estonian): "You are a compiler of an Estonian dictionary and must decide whether a register label should be added to a word or expression." The prompt used was: "Is the Estonian word 'X', in the meaning 'YY', typically used in [informal, neutral/formal][7] texts? If you cannot make the distinction or it is not clearly evident, please answer 'not applicable'. Informal texts include, for example, blogs, forums, comment sections, chat conversations, social media texts, texts with many typos, and sometimes dialogues in fiction. Please justify your choice. Base your answer solely on your training data, without using external searches or external databases (including dictionaries from EKI)[8]."

Compared to last year's prompt, this time, we added more elements, starting with specifying the meaning of a word. Since our focus was on informality, we combined neutral and formal texts into one category in the options given to the LLM, as their distinction was not relevant for this study. We also gave the LLM the option to indicate that the word does not tend to be used more in one register or the other. According to Henriksson et al. (2024), asking the models to briefly justify their decisions in preliminary tests significantly improved the results. Although we did ask the LLMs to provide justifications for their choices, we do not delve into those explanations in this paper.

The comparison of the LLMs' responses supported the hypothesis that not all words labeled as informal in the DSE are unambiguously informal. Although 892 words (67% of 1,330) were considered by all four LLMs to be used primarily in informal texts in the given meaning (e.g. *kits* 'goat' in the meaning 'complainant', *parkima* 'to park (a car)' in the meaning 'to eat a lot and greedily'), for 438 words (33%), the responses differed (e.g., *flaier* 'flyer', *kainer* 'drunk tank'). Among them, 142 were not classified as informal by any LLM (e.g., *võpsik* 'shrubs', *vaevuma* 'to bother (to do something)').

3.2.2 An evaluation benchmark

The third step that started in March was to develop an evaluation benchmark for LLMs based on these 1,330 words. Initially, we created a benchmark on a smaller set of words for a conference presentation in April.[9] However, this preliminary version is not

[7] Since our focus was on informal language, we did not consider differences in how a word in a particular meaning is used in neutral versus formal contexts.

[8] We added the part in parentheses to the prompt as the final change because Claude Sonnet 3.5 was justifying its choices based on EKI dictionaries, whereas we weren't interested in dictionary choices. That said, it cannot be ruled out that Claude nevertheless relied on EKI information and didn't make this explicit. As Davies (2025a) has said, caution is needed when interpreting LLMs' own explanations of their outputs, as their ability to accurately reflect on their decision-making processes is often no more reliable than that of humans, who are known to struggle with articulating the exact reasons behind their choices.

[9] See https://www.youtube.com/watch?v=eAUjUVUIi_s (16 September 2025).

addressed in this paper. Instead, we focus on the benchmark developed in May and June.

All kinds of methods can be used as a basis for comparison of LLMs. If we had had a dictionary where all labels were already assigned to words based on proper data analysis, then that could, in principle, have been used as a reference. However, the DSE labels hadn't yet been systematically verified. Therefore, we chose expert judgments as the basis for benchmarking.

Comparing LLMs' output with expert judgments is a method that requires clear parameters. Each word was evaluated by 3 members of the team (9 members in total), with the words divided evenly among them and their responses were blind to one another. The task was to decide whether to add the *informal* label to the word in a given meaning or not (a yes/no answer). Dictionary lookup was not allowed, and the responses from the LLMs were not visible. However, the ENC was intended to be used (as well as the corpus-based guideline that we developed based on last year's corpus analysis, see Risberg et al., 2025).

The results indicated that the team would assign the informal label to 937 words (actually, their meanings) (70.5% of 1,330 words). For 393 words (29.5%), the group would not assign the label. Comments revealed that some words were considered to be "a vivid expression," but the lexicographer wasn't sure whether that automatically made it informal (e.g., *kriimsilm* 'wolf', literally 'striped-line eye'). There were also words where the perception of what is considered "correct" by language planning disturbed objective analysis, for example, whether the "informal" *soendama* 'to warm up' really is informal, because for decades language planning has directed speakers to use the neutral word *soojendama* instead.

Out of the 1,330 words, unanimous agreement among all three annotators (3 out of 3) was reached for 589 words (44.3%). For the remaining 741 words (55.7%), two out of three annotators agreed on whether the meaning of a word is typically used in informal contexts or not. To assess the overall inter-annotator reliability, we calculated Fleiss' kappa, which was 0.189. This value indicates only slight agreement beyond chance according to commonly used interpretation scales. The relatively low agreement reflects the difficulty of the task and the inherently subjective nature of labeling informal usage in Estonian, particularly for borderline cases. While it is often recommended to use only items with full agreement (3/3) as a reliable gold standard in evaluation tasks, we opted to conduct the evaluation on the entire dataset, in order to retain a larger and more nuanced set of test items. This approach allows us to capture the complexity and uncertainty inherent in real-world lexical judgments, even though it introduces some ambiguity into reference data.

3.2.3 Benchmarking LLMs

In the fourth step of our study in June, we benchmarked LLMs using a newly created

evaluation dataset consisting of 1,330 Estonian words. To do this, we queried 12 different LLMs from various providers. While earlier stages of the study involved four models (GPT-4o, Claude Sonnet 3.5, Claude Sonnet 3.7, and Gemini 1.5 Pro), this step expanded the evaluation to include additional models such as GPT-4.5, GPT-4.1, Claude Opus 4, Claude Sonnet 4, Gemini 2.0 Flash, Gemini 2.5 Pro, Grok-3, and EuroLLM-9B (among the few open-source language models that intentionally include Estonian in their training corpus). In late August, we added the results of GPT-5 and Claude Opus 4.1. Each model's responses were collected and evaluated for accuracy based on human-annotated benchmark data. We also analyzed model accuracies separately for words where annotators were in full agreement (3 out of 3) and for those with majority agreement (2 out of 3). Table 1 presents the results of this evaluation.

Model	**Overall accuracy / Previous queries accuracy**	**100% agreement accuracy**	**Majority agreement accuracy**
GPT-4.5 preview	75.94%	84.89%	76.86%
GPT-4o	75.49% / 75.41%	84.21%	68.56%
Gemini 2.5 Pro	75.41%	84.72%	68.02%
GPT-4.1	74.66%	83.05%	67.88%
Claude Opus 4.1	74.59%	83.19%	67.75%
Claude Sonnet 3.7	74.51% / 74.14%	83.53%	67.34%
Claude Opus 4	74.29%	82.34%	67.88%
Claude Sonnet 4	74.21%	81.66%	68.29%
Gemini 2.0 Flash	73.91%	82.54%	66.94%
Claude Sonnet 3.5	73.53% / 73.53%	80.65%	67.88%
Gemini 1.5 Pro	73.46% / 73.61%	79.12%	68.96%
Grok 3	72.71%	78.95%	67.75%
GPT-5	72.71%	80.81%	66.26%
EuroLLM	49.85%	50.25%	49.53%

Table 1: Accuracy of LLMs on the benchmark

The best match between an LLM and our team's assessments was achieved by GPT-4.5 (75.94%), closely followed by GPT-4o (75.49%). In earlier tests in January, GPT-4o showed a similar level of agreement (75.41%), indicating stable performance across runs. Differences between most models were small and not statistically significant. However, EuroLLM performed significantly worse than the rest, standing out as the only model with a clearly lower level of agreement.

Across all models, accuracy was consistently higher in the 3/3 group, suggesting that the models aligned better with cases where humans had high levels of agreement. Statistical tests (z-tests for difference in proportions) confirmed that the difference in accuracy between the two groups was statistically significant ($p < 0.01$) for all models except EuroLLM. For EuroLLM, the performance was equally low in both groups (50.25% in 3/3 vs. 49.53% in 2/3), and the difference was not statistically significant.

So, do these results indicate that the top LLMs are suitable for this kind of lexicographic task? A match of only around 76% raises the question of whether LLMs can be used as one of the sources in lexicography. Based on a binomial test, we can say that the probability of achieving 76% like the top models by chance is close to zero. The overlap with human responses is statistically significantly better than random (33.3%). In other words, LLMs' responses are (statistically significantly) better than random guessing. All in all, yes, LLMs are indeed worth considering for this kind of categorization work.

3.2.4 Consistency check

Still, "the LLMs may provide wildly different 'data' for the same prompt on different occasions" (Davies, 2025b; however, human judgments can also vary when looking at the same words a second time). As a fifth step, we conducted a consistency check by comparing the latest responses of the four LLMs (GPT-4o, Claude Sonnet 3.5, Claude Sonnet 3.7, and Gemini 1.5 Pro) with their corresponding answers from February. We wanted to find out whether and how much the responses differ, and whether the difference is statistically significant.

Results for all 1,330 words showed that the responses of all four observed LLMs were highly consistent over time, with differences in accuracy of less than 0.5% between February and June runs. These small variations were not statistically significant, indicating that any observed changes are likely to be due to chance. In addition to accuracy, we analyzed the consistency of each model's responses across two runs. While most models repeated the same answer in over 98% of cases, small variations were observed – ranging from 0.6% (Gemini 1.5 Pro) to 2.5% (Claude Sonnet 3.7). Statistical tests confirmed that these variations, although minimal, are significant and indicate some degree of response instability. Nevertheless, the overall variation remains low, and based on our results, we cannot conclude that any of the LLMs became more or less

accurate or meaningfully unstable over time.

We concluded that two lexicographers using the same prompt on the same words will produce the same result with at least 97% probability. This is a difference from the use of corpus, where language data does not change. Moreover, the corpus query tool Sketch Engine gives random samples in the same order for all users. However, the interpreter interprets the data based on their own knowledge and language intuition. This is why small differences and variability are inherently part of lexicographic work, where something relatively subjective like informal language needs to be assessed.

In conclusion, the results of both experiments described in this paper are consistent and point in the same direction. In both runs of the second experiment, GPT-4o demonstrated high accuracy and stability over time, suggesting that it is currently one of the most reliable models for tasks involving labelling Estonian words informal. The minimal variation between the two runs reinforces the reproducibility of GPT-4o's outputs, which is especially important in lexicographic applications where consistency matters. GPT-4.5 (afterwards discontinued and thus unavailable for future use) outperformed others in terms of raw accuracy. However, the difference between these two models was not statistically significant. This further confirms that GPT-4o is a strong and reliable candidate for Estonian linguistic tasks, at least in categorization tasks (although it also performed well in the word meaning explanation task, see Tuulik et al., 2025). Importantly, GPT-4o also has the advantage of significantly lower costs compared to some other high-end models.

4. Discussion

According to Rundell (2012: 48), successfully completing the lexicographic task requires assurance that intuition and personal judgments are minimized, with a focus on a systematic, internally consistent method guided primarily by insights drawn from the language data. Whereas Rundell talked about using the corpus, the aim of the experiments discussed in this paper was to find out if pre-trained LLMs can support the lexicographer's decisions and reduce subjectivity by providing a more data-driven basis for adding or omitting the informal label to a word's meaning.

The results of some experiments have shown that LLMs' output is not comparable to human-quality work due to errors and hallucinations, omission, inauthenticity, "parroting" etc. (e.g., McKean & Fitzgerald, 2024; Rundell, 2024; Davies, 2025b). Although, for example, ChatGPT has performed well in assigning labels for offensive, formal, and old-fashioned English words, the non-deterministic nature of LLMs – meaning that the same question may yield different answers – implies that the results can vary when re-prompted (Trap-Jensen, 2024).

Although, from a research perspective, LLMs' output may not be reproducible, our second experiment showed that if the same prompt is used twice with the same words,

LLM responses match at least 97% of cases – a result that was much better than expected. Thus, LLMs' responses may change when re-prompted (unlike corpus data, which is always the same within a given version) this result suggests that LLMs can be used as one of the basis when deciding on the label.

While LLMs do have certain scientific limitations, there is evidence that LLMs are more effective at categorizing data than generating it (this aligns with Henriksson et al., 2024: pp. 310, whose experiment showed that LLMs perform well on annotation tasks), making them useful for determining whether words are used in informal contexts or not. For example, Davies (2025a) has shown that when analyzing corpus data, a LLM can provide register-related information. Still, LLMs have shown better results at distinguishing between clearly different genres (e.g., academic vs. fiction), whereas they struggle more with similar genres, such as fiction and film subtitles, as both involve a high degree of spoken informal language (Davies, 2025b).

Ultimately, however, the interpretation of both corpus data and LLM outputs (including corpus data analysis conducted by LLMs) remains the task of the lexicographer, who interprets the data based on their own knowledge, linguistic intuition, language experience, etc., and makes label decisions accordingly. This means that subjectivity and diversity are inherently part of this kind of work, although a lexicographer can consult additional data if needed.

All in all, a lexicographer should not ignore their linguistic intuition but rely on it to some extent, because interpreting data and making decisions based on it still requires the language experience of a native speaker. That said, there are also situations where prior knowledge and personal intuition may mislead, hinder or distract the lexicographer – for example, when evaluating word forms or meanings that language planning has long considered "incorrect" or "informal."

Henriksson et al. (2024: 313–314) concluded: "Registers that are more situationally well-defined generally map more completely to a single text type." This result relates to our observation that, for example, when talking to children, the situational register is more clearly defined than in the case of many words currently labeled as informal (i.e., colloquial). Hence, another important point to highlight is that language is not as clear-cut, nor does it always have unambiguously watertight categories[10] (see also Rundell, 2002) as lexicography would require, for example, when deciding whether to add an informal label to a word meaning or not, or language planning would require, where they prescriptively determine whether a word, word form, or meaning is "correct" or not (for a critique of the prescriptive approach, see Pullum, 2023).

Accordingly, one issue related to dictionaries is that users tend to interpret even

[10] There are many word meanings that do not fall strictly into informal language (see Section 2.1) and registers themselves can be recognized by certain linguistic markers or situational contexts, but they cannot always be clearly defined based on them (Henriksson et al., 2024: 308).

descriptive dictionaries as normative. Labels, too, have often been seen as normative information (Langemets et al., 2024: 753; Trap-Jensen, 2002). While this has previously been the case with the DSE, in the compilation of the new DSE (2025), most labels are treated as descriptive information. That said, it cannot be ruled out that some prescriptive labels may remain, as the printed dictionary must be published by the end of 2025 and cannot be updated for another seven years and human error can occur when making decisions. In contrast, labels in the web-based *EKI Combined Dictionary* can be revised later if needed.

However, thanks to the compilation of the benchmark for evaluating the LLMs and analyzing word usage, the compilers of the DSE 2025 removed the informal label from some word meanings where the label was unnecessary (e.g., *ribi* 'rib', *elektrikontakt* 'power outlet', *kirvereegel* 'rule of thumb'). In February, there were 1,330 word meanings with the informal label, whereas by September the number had decreased to 1,202.

5. Future perspectives

Although there have been doubts about whether LLMs will eventually develop to the point where dictionaries and lexicography are no longer needed (see de Schryver, 2023), lexicography continues to play an important role – particularly for underrepresented languages where LLMs still face limitations (Lew, 2024). After all, user experience studies have shown that one of the most important features of a dictionary is the reliability of its content (Müller-Spitzer & Koplenig, 2014: 168; Langemets et al., 2024). Users need confidence that the information in the dictionary is curated, i.e., it "has been selected to reflect what is most characteristic of the way a given word behaves" (Rundell, 2024). Still, "the role of lexicographers is likely to shift towards guiding and refining increasingly automated tools, ensuring ethical linguistic data use, and counteracting AI biases" (Lew, 2024).

The experiments described in this paper provide a basis for future improvements, as we used pre-trained LLMs and the corpus separately. Relying solely on the training data of LLMs is problematic, as there are cases where they simply "parrot" information from published sources, e.g., dictionaries (Davies, 2025b; this was also evident in Claude Sonnet 3.5's justifications during the early tests of our second experiment). At times, such external sources include prescriptive judgments that are not grounded in actual language use, but the LLM is not aware of this. Corpus data can be used to support and correct its output. (Davies, 2025b) Therefore, a future direction is to have LLMs analyze and categorize corpus data (as in Davies, 2025a), which could lead to even more objective insights into how words and their meanings are used across different registers.

LLMs should be efficient and easy to use for lexicographers, as time is of the essence when compiling a dictionary. Nevertheless, not every lexicographer is a skilled programmer. Therefore, while so far, we have conducted experiments through the API,

in the future we should also explore the possibilities of using the chat interfaces of the LLMs. These interfaces do not require programming skills, which makes them more accessible, but they also come with notable drawbacks, such as unwanted wordiness, a tendency to please the user, reduced reproducibility of results, and limited scalability for large datasets.

We continue to examine what kind of data, and in which cases, support the claim that a word in one of its meanings tends to be used primarily in informal contexts and thus adding an informal label to that meaning in the dictionary is justified. Until then, however, we can say that we are moving towards even more data-driven decisions, using both corpus analysis and LLMs as analytical tools.

As a dictionary should serve "the needs of the greatest number of users in the greatest number of situations" (Rundell, 2002: 150), it ought to offer data-driven generalizations about word usage and indicate when a word, in one of its meanings, tends to be used in contexts that deviate from neutral (or formal) register. Since determining registers also requires actual usage context, in August we launched another experiment to examine the extent to which corpus data can help LLMs detect the registers of word meanings. As of early October, this experiment is ongoing, but it has already shown that analyzing corpus data offers advantages over relying solely on LLMs' training data (for example, in identifying meanings and assigning them a register).

6. Acknowledgements

This paper was funded by R&D project "Applying large language models to lexicography: new opportunities and challenges" (project executor: Institute of the Estonian Language; funder: Estonian Ministry of Education and Research). We are also grateful to Sirli Zupping for her help with data analysis.

7. References

Anthropic. (2024). Claude [Large Language Model]. https://www.anthropic.com/claude. https://www.anthropic.com/claude (19(16 September 2025)

Baayen, R.H. (2024). The wompom. *Corpus Linguistics and Linguistic Theory*, 20(3), pp. 615–648.

Beal, J., Lukač, M. & Straaijer, R. (2023). *The Routledge Handbook of Linguistic Prescriptivism.* London, New York: Routledge.

Bender, E.M., Gebru, T., McMillan-Major, A. & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*. New York, NY, USA: Association for Computing Machinery, 610–623.

Bi, X., Chen, D., Chen, G., Chen, S., Dai, D., Deng, C., Ding, H., Dong, K., Du, Q., Fu, Z. and Gao, H., et al. (2024). DeepSeek LLM: Scaling open-source language

models with longtermism. *arXiv preprint arXiv:2401.02954*. (16 September 2025)

Biber, D. & Conrad, S. (2009). *Register, Genre, and Style (Cambridge Textbooks in Linguistics)*. Cambridge: Cambridge University Press.

Biber, D. & Egbert, J. (2023). What is a register? Accounting for linguistic and situational variation within – and outside of – textual varieties. *Register Studies*, 5. https://www.jbe-platform.com/content/journals/10.1075/rs.00004.bib. (16 September 2025)

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A. & et al. (2020). Language Models are Few-Shot Learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan & H. Lin (eds.). *Advances in Neural Information Processing Systems*, 33, pp. 1877–1901. Curran Associates, Inc.

Claude Sonnet 3 = Anthropic (2024). The Claude 3 Model Family: Opus, Sonnet, Haiku. https://www.anthropic.com/news/claude-3-family. (16 September 2025)

Claude Sonnet 3.5 = Anthorpic. (2024). Claude 3.5 Sonnet. https://www.anthropic.com/news/claude-3-5-sonnet. (16 September 2025)

Claude Sonnet 3.7 = Anthropic. (2025). Claude 3.7 Sonnet and Claude Code. https://www.anthropic.com/news/claude-3-7-sonnet. (16 September 2025)

Claude Sonnet 4 = Anthorpic. (2025). Introducing Claude 4. https://www.anthropic.com/news/claude-4. (16 September 2025)

Claude Opus 4 = Anthorpic. (2025). Introducing Claude 4. https://www.anthropic.com/news/claude-4. (16 September 2025)

Claude Opus 4.1 = Anthropic. (2025). System Card Addendum: Claude Opus 4.1. https://www.anthropic.com/claude-opus-4-1-system-card. (16 September 2025)

Davies, M. (2025a). Comparing and Integrating Information from Corpora and AI/LLMS. *EURALEX Talks*, 16 April 2025. https://videolectures.net/videos/EURALEXTalks_davies_information. (16 September 2025)

Davies, M. (2025b). Integrating AI / LLMs into English-Corpora.org. https://www.english-corpora.org/ai-llms. (16 September 2025)

De Schryver, G.-M. (2023). Generative AI and Lexicography: The Current State of the Art Using ChatGPT. *International Journal of Lexicography*, 36(4), pp. 355–387.

DSE 2018 = Eesti õigekeelsussõnaraamat ÕS 2018 [Dictionary of Standard Estonian]. T. Erelt, T. Leemets, S. Mäearu & M. Raadik (eds.). Eesti Keele Instituut. Tallinn: EKSA.

EKI Combined Dictionary = *Eesti Keele Instituudi ühendsõnastik 2025.* Eesti Keele Instituut, Sõnaveeb. https://sonaveeb.ee. (16 September 2025)

EuroLLM = Martins, P.H., Alves, J., Fernandes, P., Guerreiro, N.M., Rei, R., Farajian, A., Klimaszewski, M., Alves, D.M., Pombal, J., Faysse, M. and Colombo, P. (2025) *EuroLLM-9B: Technical Report*. arXiv preprint arXiv:2506.04079. (16 September 2025)

Gemini Team, Google. (2023). Gemini: A Family of Highly Capable Multimodal Models. arXiv preprint arXiv:2312.11805 (16 September 2025)

Gemini 1.5 Pro = Gemini Team, Google. (2024). *Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context.* arXiv preprint arXiv:2403.05530 (16 September 2025)

GPT-4 = Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.-L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S. & et al. (2023). *Gpt-4 technical report.* arXiv preprint arXiv:2303.08774 (16 September 2025)

GPT-4.1 = OpenAI. (2025). Introducing GPT-4.1 in the API. https://openai.com/index/gpt-4-1/. (16 September 2025)

GPT-4.5 = OpenAI. (2025). OpenAI GPT-4.5 System Card. https://openai.com/index/gpt-4-5-system-card/. (16 September 2025)

GPT-4o = Hurst, A., Lerer, A., Goucher, A.-P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A.J. & et al. (2024). GPT-4o system card. arXiv preprint arXiv:2410.21276 (16 September 2025)

GPT-5 = OpenAI. (2025). GPT-5 System Card. https://cdn.openai.com/gpt-5-system-card.pdf. (16 September 2025)

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., et al. (2024). *The llama 3 herd of models. arXiv preprint arXiv:2407.21783.* (16 September 2025)

Grok 3 =. xAI. (2025). Grok 3 Beta — The Age of Reasoning Agents. https://x.ai/news/grok-3. (16 September 2025)

Hadi, M.-U., Qureshi, R., Shah, A., Irfan, M., Zafar, A., Shaikh, M.B., Akhtar, N., Wu, J., Mirjalili, S & et al. (2023). *A survey on large language models: Applications, challenges, limitations, and practical usage.* Authorea Preprints. Authorea.

Henriksson, E., Myntti, A., Hellström, S., Erten-Johansson, S., Eskelinen, A., Repo, L. & Laippala, V. (2024). From Discrete to Continuous Classes: A Situational Analysis of Multilingual Web Registers with LLM Annotations. *Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities*, pp. 308–318.

Jakubíček, M. & Rundell, M. (2023). The end of lexicography? Can ChatGPT outperform current tools for post-editing lexicography? *Electronic Lexicography in the 21st Century: Invisible Lexicography, Brno, Czechia. eLex 2023*, pp. 518–533.

Karelson, R. (1990). "Eesti kirjakeele seletussõnaraamat" tegija pilgu läbi. *Keel ja Kirjandus*, 1, pp. 24–34.

Kasik, R. (2021). Normikeel ja ühiskeel – eesti keel. *Sirp* 23 July.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P. & Suchomel, V. (2014). The Sketch Engine: ten years on. *Lexicography*, 1(1), pp. 7–36.

Klosa-Kückelhaus, A. & Tiberius, C. (2024). The Lexicographic Process Revisited. *International Journal of Lexicography*, 38(1), pp. 1–12.

Koppel, K. & Kallas, J. (2022). Eesti keele ühendkorpuste sari 2013–2021: mahukaim eestikeelsete digitekstide kogu. *Eesti Rakenduslingvistika Ühingu aastaraamat*, 18, pp. 207–228.

Koppel, K., Tavast, A., Langemets, M., & Kallas, J. (2019). Aggregating dictionaries into the language portal *Sõnaveeb*: Issues with and without a solution. *Electronic lexicography in the 21st century. Proceedings of the eLex 2019 conference. 1-3 October 2019, Sintra, Portugal. Lexical Computing*, pp. 1–3.

Koppel, K., Kallas, J., Jürviste, M. & Kaljumäe, H. (2023). *Estonian National Corpus 2023*. Lexical Computing Ltd. / Eesti Keele Instituut.

Langemets, M., Risberg, L. & Algvere, K. (2024). To Dream or Not to Dream About 'Correct' Meanings? Insights into the User Experience Survey. In *XXI EURALEX International Congress. Cavtat, Croatia*, 741−760.

Lew, R. (2024). Dictionaries and lexicography in the AI era. *Humanities and Social Sciences Communications volume* 11. https://www.nature.com/articles/s41599-024-02889-7. (16 September 2025)

Li, Z., Shi, Y., Liu, Z., Yang, F., Liu, N. & Du, M. (2024). Quantifying multilingual performance of large language models across languages. arXiv preprint arXiv:2404.11553. (16 September 2025)

Marcondes, F.S., Adelino de C. O. S. Gala, Manuel Rodrigues, José João Almeida & Paulo Novais. (2024). Lexicon Annotation with LLM: A Proof of Concept with ChatGPT. *International Conference on Hybrid Artificial Intelligence Systems* (Lecture Notes in Computer Science), pp. 190–200. https://link.springer.com/chapter/10.1007/978-3-031-74186-9_16. (16 September 2025)

McKean, E. & Fitzgerald, W. (2024). The ROI of AI in lexicography. *Lexicography* 11(1). https://utppublishing.com/doi/abs/10.1558/lexi.27569. (16 September 2025)

Müller-Spitzer, C. & Koplenig, A. (2014). Online dictionaries: expectations and demands. In C. Müller-Spitzer (ed.). *Using Online Dictionaries. (Lexicographica. Series Maior 145.)* Walter de Gruyter, pp. 143–188.

Nguyen, X.-P., Aljunied, M., Joty, S. & Bing, L. (2024). Democratizing LLMs for Low-Resource Languages by Leveraging their English Dominant Abilities with Linguistically-Diverse Prompts. In L.-W. Ku, A. Martins & V. Srikumar (eds.). *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*. Bangkok, Thailand: Association for Computational Linguistics, pp. 3501–3516.

OpenAI. (2023). GPT-4 Technical Report. https://openai.com/research/gpt-4. (16 September 2025)

Pajusalu, R. (2009). *Sõna ja tähendus*. Tallinn: Eesti Keele Sihtasutus.

Pullum, G.K. (2023). Why grammars have to be normative – and prescriptivists have to be scientific. In Beal, J., Lukač, M. & Straaijer, R. (2023). *The Routledge Handbook of Linguistic Prescriptivism*. London, New York: Routledge, pp. 3–16.

Raag, R. (2008). *Talurahva keelest riigikeeleks*. Tartu: AS Atlex.

Regular Evaluation Report 2024. Arts and Humanities. https://etag.ee/wp-content/uploads/2025/05/Eesti-Keele-Instituut.pdf. (16 September 2025)

Risberg, L. (2024). *Sõnatähendused ja sõnaraamat. Kasutuspõhine sisend eesti keelekorraldusele.* (Dissertationes philologiae Estonicae Universitatis Tartuensis 52.) Tartu: Tartu Ülikooli Kirjastus.

Risberg, L., Tuulik, M., Langemets, M., Koppel, K., Vainik, E., Prangel, E. & Aedmaa, E. (2025). Keelekorpus kui leksikograafi abiline kõnekeelsuse tuvastamisel [Using corpus data to support lexicographers in identifying informal language]. *Keel ja Kirjandus* 7, pp. 605–624. https://doi.org/10.54013/kk811a3. (16 September 2025)

Rotter, S. & Liu, M. (2023). Interlocutor relation predicts the formality of the conversation: An experiment in American and British English. *Register Aspects of Language in Situation (REALIS)* 2(2), pp. 1–27. https://doi.org/10.18452/26192. (16 September 2025)

Rundell, M. (2002). Good old-fashioned lexicography: Human judgement and the limits of automation. In M.-H. Corréard (ed.). *Lexicography and Natural Language Processing: A Festschrift in honour of B. T. S. Atkins. Stuttgart: EURALEX*, pp. 138–155.

Rundell, M. (2012). It works in practice but will it work in theory? The uneasy relationship between lexicography and matters theoretical (Hornby Lecture). In Fjeld, R.V. & Torjusen, J.M. (eds.). *Proceedings of the 15th EURALEX Congress. Oslo: University of Oslo*, pp. 47–92.

Rundell, M. (2024). Automating the Creation of Dictionaries: Are We Nearly There? – Humanising *Language Teaching*, 26(1). https://www.hltmag.co.uk/feb24/automating-the-creation-of-dictionaries. (16 September 2025)

Sõnaveeb. Language portal, Institute of the Estonian Language. https://sonaveeb.ee. (16 September 2025)

Tao, Y., Viberg, O., Baker, R.S. & Kizilcec, R.F. (2024). Cultural bias and cultural alignment of large language models. In M. Muthukrishna (ed.). *PNAS Nexus*, 3(9).

Trap-Jensen, L. (2002). Descriptive and Normative Aspects of Lexicographic Decision-Making: The Borderline Cases. In *Proceedings of the Tenth EURALEX International Congress. Copenhagen*, pp. 503–509.

Trap-Jensen, L. (2024). The Best of Two Worlds: Exploring the Synergy between Human Expertise and AI in Lexicography. https://lexicography21.iliauni.edu.ge/wp-content/uploads/2024/06/03_Lars-Trap-Jensen.pdf. (16 September 2025)

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F. and Rodriguez, A. (2023*). Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971.* (16 September 2025)

Tu, N.D.T., Lang, C. & Brunner, A. (2025). LLM fails. Gescheiterte Experimente mit Generativer KI und was wir daraus lernen können. *Workshop am 8. und 9. April 2025, Leibniz-Institut für Deutsche Sprache*. https://www.ids-

mannheim.de/home/lexiktagungen/llm-fails/. (16 September 2025)
Tuulik, M., Vainik, E., Prangel, E., Langemets, M., Aedmaa, E., Koppel, K. & Risberg, L. (2025). Tähenduste seletamine leksikograafias: kuivõrd on abi suurtest keelemudelitest? [Describing senses for lexicography: how helpful are large language models?] *Eesti ja soome-ugri keeleteaduse ajakiri. Journal of Estonian and Finno-Ugric Linguistics*, 16(2), pp. 147–176. https://doi.org/10.12697/jeful.2025.16.2.05. (Available in 6–10 October 2025)
Vaik, K. (2024). *Beyond Genres: A Dimensional Text Model for Text Classification.* (Dissertationes linguisticae Universitatis Tartuensis 47.) Tartu: Tartu Ülikooli Kirjastus.
Vare, S. (2001). Üldkeele ja oskuskeele nihestunud suhe. *Keel ja Kirjandus*, 7, pp. 455–472.
Yu, X., Zhang, Z., Niu, F., Hu, X., Xia, X. & Grundy, J. (2024). What Makes a High-Quality Training Dataset for Large Language Models: A Practitioners' Perspective. In *Proceedings of the 39th IEEE/ACM International Conference on Automated Software Engineering.* New York, NY, USA: Association for Computing Machinery, 656–668.