Can advances in NLP lead to worse results for Uralic languages and how can we fight back? Experiences from the world of automatic spell-checking and correction for Finnish

Flammie A Pirinen
Divvun
UiT—Norgga árktalaš universitehta
Tromsø, Norway
first.last@uit.no

Abstract

Spell-checking and correction is a ubiquitous application within text input in modern technology, and in some ways or another, if you type texts on a keyboard or a mobile phone, there will probably be an underlying spelling corrector running. The spell checkers have been around for decades, initially based on dictionaries and grammar rules, nowadays increasingly based on statistical data or large language models. In recent years, however, there has been a growing concern about the quality of these modern spell-checkers. In this article, we show that the spell-checkers for Finnish have gotten significantly worse in their modern implementations compared to their traditional knowledge-driven versions. We propose that this can have critical consequences for the quality of texts produced, as well as literacy overall. We furthermore speculate if it would be possible to get spell-checking and correction back on track for Uralic languages in modern systems.

1 Introduction

Spell-checking and correction is a quintessential *natural language processing* (NLP) task. It has been part of the NLP ecosystem for decades now, from the very early days of processing texts with computers. It has become so ubiquitous that it exists in most text editing products without users even paying much attention to it, and it has been viewed as somewhat of a solved problem within the scientific study for the last few decades. While there has not been much focus on spell-checking and correction in recent years, we as linguists have noticed something quite problematic with the contemporary systems. Namely, we have noticed a drop in quality of writing in our native languages on the Internet discussion forums. This is increasingly shown in the frustrations by the native writers: "autocorrect wrote it, and it is too hard to fix it by hand". On this basis, we set out to study, if the contemporary spell-checking and correction systems have become worse in our language. Our hypothesis is, that modern autocorrecting spell-checking and correction systems are based on data-driven methods and lately large language models, which may work adequately with English—not the least because over 90 % of the training data is in English[1]—but which actually fail to recognise words of non-English languages with potentially more complicated morphology.

Our *research question* in this short paper is, are data-driven and large language model based spelling checkers and correctors worse than traditional knowledge-based ones? Our initial hypothesis, based on everyday observations, is that spell-checking tools have gotten significantly worse in the past few decades, in pace with the introduction of data-driven and 'AI'-driven models. We study the spell-checking and correction results by three popular systems for *Finnish*.

2 Background

There is a long history of spell-checking and correction in language technology, starting from early days of SPELL, a spell-checker based on a dictionary or a word-list and few simple rules to modify suffixes. Earnest (1976) places initial use of their spelling correction to 1969. This system's descendants—ispell, aspell and hunspell and so forth—have been in use in some of the most popular browsers and office suites up to the 2000s. There have been several comprehensive scientific surveys of spell-checking and correction, for example Kukich (1992). As of last few decades, office suites have started using built-in, closed-source, statistical spell-checkers and more recently, overarching AI assistants which also do spell-checking,

[1] c.f. e.g. https://github.com/openai/gpt-3/blob/master/dataset_statistics/languages_by_word_count.csv

Proceedings of the 10th International Workshop on Computational Linguistics for Uralic Languages, pages 22–28

a similar development is happening in browsers, mobile phones and operating systems. One of the most influential initial works on data-driven spell-checking and correction is Google's Norvig's spelling corrector (Norvig, 2009). Technical details of the most modern commercial spelling correctors are not openly documented as far as we know.

When researching on existing studies on LLM-based spelling correctors and especially comparisons between LLM and traditional methods, one topic that dominates the results is spell-checking and correction for students / L2 / EFL users of English (Jaashan and Alashabi, 2025; Gayed et al., 2022). A gap in research we hope to address with this experiment and its followups, is, therefore, a comparative study, and for L1 users, in non-English.

In our work, we build NLP tools and software, mainly targeting less-resourced, minority and Indigenous languages, but we also create tools that are language agnostic and usable for all. Spell-checking and correction and related software is a key tool for digital language survival for minority languages, and it is also increasingly important for ever larger and more majority languages, apart from the largest few. This has come to be also, because the contemporary data-driven language technology is strongly based on big data, that has been written by humans in correctly spelled and grammatical language.

In this experiment, we have chosen to use Finnish. Finnish is a national, majority language in Finland, it is not low-resource by any stretch of the imagination. We estimate it is likely to be in the top 50 of the most resourceful languages in the world. While we are more interested in low-resource languages and settings, having moderately resourced Uralic language works well for our initial experimentation. We have existing resources such as corpora and established automatic spell-checkers, which we might not find on lower resourced languages. There is also existing research on state of the Finnish NLP (Hämäläinen and Alnajjar, 2021) including spell-checking and correction. Furthermore, Finnish is not an Indo-European language, and has a slightly more complicated morphology than most IE languages, which makes it more comparable towards many of the minority and under-resourced languages relevant to our research. Finally, we have native speakers of Finnish, which in our opinion is critical in doing meaningful qualitative studies on language technology software; without linguistic error analysis and human interpretation of the results, it is impossible to make meaningful explanation of how useful or harmful the underlying system is for actual end users.

Linguistically, Finnish is a Uralic language with some 5 million speakers, mainly in Finland. Morphologically, Finnish has what we call slightly more complex morphology, in terms of what matters for spell-checking and correction this means that there are on average thousands of word-forms per word, instead of around 5 like in English or few dozens like in most IE languages. Finnish also has productive compounding, which means you can put two word-forms together without a space to create a new word, that does not necessarily exist in the dictionary, on the fly. Finnish has had a literary culture for several hundreds of years and has a strong nationally backed standardisation body, is a primary language in schools and in public. It is also a majority language for several Indigenous and minority languages, which is one of the motivations for us to work on it as well.

3 Methods and experimental setup

In this work, we compare and contrast spell-checking and correction from the end-user point of view. We test three different systems: one based on knowledge-driven paradigm and two based on data-driven approach. The knowledge-driven spell-checker is an open source, rule-based product, whereas the data-driven products we experiment with are commercial and closed-source.

The rule-based spell-checking and correction is a freely available open source implementation of Finnish spell-checking found on the GitHub called omorfi[2], their implementation is based on finite-state spell-checking (Pirinen and Lindén, 2014). This spell-checker uses an underlying dictionary and morphological rules to recognise valid word-forms without context, and uses finite-state error modelling technology to create suggestions for corrections.

We use Google's spell-checking and correction as a black-box, we have not found technical documentation detailing it, but we estimate that it is at least in part based on statistical methods and or large language models based on the company's recent focuses and public statements.[3] The

[2] https://github.com/flammie/omorfi/

[3] Searching online leads to old posts like: https://work

function in Google Docs interface is found under spelling and grammar checking, we have crossed off grammar-checking and only included spelling.

For a product that is certainly using large language models, we test ChatGPT (OpenAI, 2025), and we use it as a black box with the version available to us via our university account. We use the web-based ChatGPT user interface to query spelling corrections from the language model via its natural language user interface in the same way an average end user likely would.

The experiments have been performed in May 2025, some details are included in the appendix A, but since they are closed commercial products, we do not expect to be able to have reproducible results with them in any case.

4 Data

To test the spell-checking and correction we have used a Finnish translation of *Alice's Adventures in Wonderland* from Project Gutenberg[4] which is in public domain. This book is a fantasy novel aimed for children, and contains creative use of language which makes it very suitable for natural language processing testing. The translation has been made in early 20th century which matches the most modern standard written Finnish with almost no deviations. In general, proofreading at the times of the publication was highly valued and efficient, and we expect the manuscript to be mostly error-free barring potential mistakes in gutenberg's encoding. The non-word errors we have found and verified are listed in the error analysis section 5.1. The book consists of 18,861 space-separated tokens (after removing project Gutenberg's licence, preamble and postamble).

5 Results

To measure the spelling error correctors, we went through all the words that were flagged as spelling errors, and categorised them into two categories: *false positives*, where a correctly spelled word was flagged as incorrect, and *true positives*, where the flagged word did contain a spelling error. This was done by a native speaker who had access to the error in context, even though the decision was made solely on whether the word is a valid word in the language at all or not (i.e. it can also be decided without context as traditional non-word spelling corrector does). The breakdown of errors and flaggings is shown in the Table **??**, we also provide a calculations of precision, recall and $F_{0.5}$, the parametre 0.5 for β is selected since our starting point is that false positives are more critical problem in spell-checking than false negatives.

Error \ System	**Google**	**ChatGPT***	**omorfi**
False Positive	565	75	59
False Negative	22	59	20
True Positive	41	4	43
Precision	0.07	0.05	0.42
Recall	0.65	0.06	0.68
F-Score ($F_{0.5}$)	0.08	0.05	**0.46**

Table 1: Quantitative evaluation of error types by systems. * ChatGPT results are not proportional due to reasons explained in the chapter. For main findings, read the qualitative error analysis.

5.1 Error analysis

We have further categorised the errors flagged by the spelling correctors into error types, based on linguistic insight and world knowledge. We hypothesise this will help give an impression of the impact these errors have on the user experience, this impact is further discussed in the section 6 below. The summary of errors is given in table 2, some of the error classes are not mutually exclusive and the numbers in the rows do not add up to the total.

One of the largest groups of false positives in all systems' data is compound words, particularly the types that do not appear in dictionary: for Google's spell-checking **compound** nouns like *pääkallonkuva* (picture of a skull) or *kyynellammikko* (lake of tears) were consistently underlined, for ChatGPT we have e.g. *herttakuningatar* (queen of hearts) and for omorfi we saw compound adverbs like *tuulennopeasti* (in wind's speed). From **derivational** forms, all systems stumbled on *ruukkusen* (little jar's∼jarful[?]). Some of the false positives found by Google can also be described as being part of complex morphology that

space.google.com/blog/productivity-collaboration/everyday-ai-beyond-spell-check-how-google-docs-is-smart-enough-to-correct-grammar, but we cannot know for sure if this kind of information is up-to-date.

[4] https://www.gutenberg.org/ebooks/46569 for reproducibility we have the version we used in our GitHub at https://github.com/flammie/purplemonkeydishwasher/tree/master/2025-iwclul/reprodata; this is also for access from within Germany, Italy or other countries with extreme copyright restrictions where Project Gutenberg may not be available.

is a bit half-ways between *inflectional* and derivational morphology, for example *myöhästynkin* (I will be late too), *elämäniässään* (in their lifetime), *vaikeroidessaan* (while they were whining), that is, enclitic particles, possessive suffixes and non-finite verb forms in combinations that—in all likelihood have not been many times in sufficiently large corpora—throw Google's spelling checker off the track. The commonality for errors in this category is that there are at least two distinct inflectional suffixes in the word-form. Perhaps surprisingly, also **proper nouns** show up as false positives, even though traditionally maybe it has been common practice to ignore titlecased words: Google finds *Ellakaan* (Ella too) and *Vilhelmiä* (of Vilhelm) errors, and omorfi finds *Morcar* and *Stigand*. The classes as laid out in the table 2 are not mutually exclusive, i.e. a **compound** form can also have a **derivation** and a **proper noun** can have a **inflectional** possessive suffix, in these cases we have simply counted the error in both classes. To illustrate the overlapping between categories, for example *Irvikissakaan* (Cheshire cat neither, lit. grinning-y cat) is a proper noun compound with inflectional ending. There are handful of words that do not seem to fall into any categories; for omorfi we can simply note they are missing from the dictionary, e.g. *satakaunoja* (an old word for some flower) or *siekailuun* (into scrupulousness) whereas with data-driven models we can assume the words themselves are so rare that they do not show up enough in the training materials, e.g. *pulppusivat* (bubbled up) or *pulikoinut* (drudged about), but there are some that are even harder to diagnose, such as *nurmen* (grass') and *vai* (or).

The true positives in the text fall into following categories: unexpected hyphenation caused by creative language use (recreation of typeset poems: *tar-kemmin* (tarkemmin), *päi-villä* (päivillä), and *veruk-keella* (verukkeella)), lengthening of letters for emphasis (*li-iemi* (liemi), *ku-ulta* (kulta) and *ihana-ainen* (ihanainen)), foreign words (*Oú*, *est*, and *chatte*), dialectal, informal or poetic forms (*teälhän* (täällähän), *käshän* (käsihän), *näkkyy* (näkyy), *käs* (käsi), *sittennii* (sitenkin), *pyssyy* (pyssyä), *ruppee* (rupeaa), *pentus* (pentusi), *juur* (juuri), *loitoll'* (loitolla), *täss'* (tässä), *kuus* (kuusi), *tavaraks* (tavaraksi), *niill'* (niillä), and *tuoss*), compounding mistakes (*mitenpäin* (miten päin), *missäpäin* (missä päin), *käsikädessä* (käsi kädessä), *sukkajalassa* (sukka jalassa), *ranskankieltä*, *tipo (tiessään)* (tipotiessään, a non-word error since tipo by itself is not a dictionary word but a reduplicative form), which old standard may have allowed), old forms (*sebraa* (seepraa), *merikilpiö* (?merikilpikonna) again permissible by older standards) onomatopoeia (*liuskis*, *läyskis*) and two typoes (*antipatiioiksi* (antipatioiksi) and *purstölleni* (pyrstölleni). We consider all of these non-words (and eventually true positives) since it is expected for a typical spell-checker to flag them, even though not all of these need to be fixed in context of this book.

Error \ System	**Google**	**ChatGPT**	**omorfi**
Compound	169	38	18
Derivation	21	9	7
Inflection	211	6	4
Proper noun	12	0	8
Other	171	24	32
Total	611	70⋆	57

Table 2: Error analysis of false positives in Alice in Wonderland by three systems. Classes are not mutually exclusive and may not add up to totals per column. ⋆ChatGPT started to give empty answers and repeat from the beginning after 70 spelling errors.

6 Discussion

While we expected to find some false positives from all the methods, we were quite surprised indeed to discover how many false positives Google's spelling error correction flags: over 600 errors in a book of 70 pages means that you see several wrong red squiggly lines on every page. This would have been unacceptable and catastrophical for an office suite in the 1990s, it is alarming that this is not the case any more. The fact that this is given to end users without warnings is starting to be borderline ethically questionable, it has a real possibility to be destructive to language and culture, as many of the false positives concern morphologically complexer forms will contribute to make the language poorer, as language learners and less confident writers will surely follow the advice of spelling correction program.

ChatGPT's spell-checking is interesting since, despite the fact that we specifically asked it to only include non-words, kept including real-word errors. ChatGPT also includes a helpful explanation for each spelling error it discovers, this is the opposite of Google doc's system which only provides

a single correction suggestion without any background. Unfortunately, the explanation often ends up being nonsensical, for example:

ChatGPT

"torkuksissa - This word does not exist in Finnish. Likely a typo for "torkuksissa" (a colloquial form of "torkuksissa")."

it reminds us in form the kind of reasonable advice you would get from a helpful grammar corrector, but content is absolutely mind-boggling and in fact gas-lighting.

The rule-based spell-checkers also only give very limited feedback to the end-user, a squiggly red underline to communicate that the word is not in the dictionary and a list of most common words within a few mistaken keystrokes away. Sometimes rule-based spell-checkers are used as a part of a grammatical error correction system where the grammar-checker can provide context, but it is typically a very mechanical and limited explanation. Perhaps an ideal hybrid system could be to harness ChatGPT's power to create user-friendly descriptions in addition to rule-based knowledge of actual dictionary and grammar, in style of this actual example from ChatGPT:

ChatGPT

"herttuatar - While valid, it is an older term (archaic) for "duchess.""

In this case, ChatGPT had flagged a common word as archaic, but it still gives the end user information based on which they can more confidently ignore the suggestion and not left feeling confused or annoyed. Certainly one could argue that if it was a modern text about Finnish society and not a translated text of older times, there would be much less talk about duchesses.

The correction mechanism in Google Docs only gives out one suggestion for corrections, this leads to many cases where it often ends up actually suggesting the mistake that users commonly make, exactly the opposite of what we would want from a spelling corrector. This happens for example for replacing forms of word *koettaa* (attempt) to word *koittaa* (dawn, verb of sun/morning), a very common mistake that beginner writers make. It also suggests to split compound words, and on one occasion it wants to replace *ja pani* (and put) with *japani* (Japanese).

We are concerned that the lowered quality of spell-checking that is included in all of our devices and office suites ultimately contributes to lower quality of texts and literacy, and while the effect is already noticeable for majority languages like Finnish, the effect will be even greater for less resourced, more minoritised and Indigenous languages. Some experts have speculated that the aggressive push for AI-based writing aids into both office suites and also in the mobile phone platforms will eventually lead into removal of traditional and alternative spell-checkers in these contexts; if this happens with the spell-checkers such as current spell-checker of Google Docs, it will spell a disaster for Finnish language literacy.

7 Conclusion

In this article, we have shown through experimental means that data-driven spell-checking and correction is much worse for Finnish language than the traditional rule-based approaches. Nevertheless, the main systems provided for spell-checking and correction in many contemporary contexts are using this kind of spelling correctors for Finnish, without any easy way to change them.

Limitations

In this article, we have performed an experiment for one language and one book, based on limitations of time and human resources: judging and manually analysing spelling error corrections requires full read-through of the whole text by a person with native-like language skills who has been trained in proofreading. There is ample anecdotal evidence that spell-checkers underperform for other Uralic and minority languages that can be discovered by simple search into language learning communities in discussion forums like reddit. More research on other languages is needed, and we hope our work gives inspiration for other researchers.

The experiments on large language models have been made on commercial systems, which makes reproducibility virtually impossible. Furthermore the version of ChatGPT we had an access to did not manage to error check the whole text correctly, for future revisions we will try to find an alternative that can be more functional; anyways this highlights the problems that average end-user will face trying to spell-check their texts the way that is available to them. Training and fine-tuning our own

model would not have been a realistic evaluation setup for the purposes of this article.

Ethics

The experiments and analysis have been made by fully paid colleagues, no underpaid crowd-workers have been hired for this experiment. The LLMs used in the experiment waste unethically large amounts of energy and water, while we have tried to minimise the wastage, our aim for this article is to curb unnecessary overuse of LLM-based systems through which we hope to achieve a net positive.

References

Les Earnest. 1976. A look back at an office of the future. In *IIASA PROCEEDINGS SERIES*, page 119.

John Maurice Gayed, May Kristine Jonson Carlon, Angelu Mari Oriola, and Jeffrey S Cross. 2022. Exploring an ai-based writing assistant's impact on english language learners. *Computers and Education: Artificial Intelligence*, 3:100055.

Mika Hämäläinen and Khalid Alnajjar. 2021. The current state of finnish NLP. *CoRR*, abs/2109.11326.

Hasan Mohammed Saleh Jaashan and Abdulazziz Ali Alashabi. 2025. Using ai large language model (llm-chatgpt) to mitigate spelling errors of efl learners. In *Forum for Linguistic Studies*, volume 7, pages 328–339.

Karen Kukich. 1992. Techniques for automatically correcting words in text. *ACM computing surveys (CSUR)*, 24(4):377–439.

Peter Norvig. 2009. Natural language corpus data. *Beautiful data*, pages 219–242.

OpenAI. 2025. Chat-gpt 4o. Online, accessed 2025-05.

Tommi A Pirinen and Krister Lindén. 2014. State-of-the-art in weighted finite-state spell-checking. In *Computational Linguistics and Intelligent Text Processing: 15th International Conference, CICLing 2014, Kathmandu, Nepal, April 6-12, 2014, Proceedings, Part II 15*, pages 519–532. Springer.

A Versions and parametres

The ChatGPT we use identifies itself as ChatGPT-4o. ChatGPT prompt is in figure 1.[5] Omorfi is version 0.9.11[6] Google Docs does not seem to have version identification available in the usual places, we used in 2025-05.[7]

[5] `https:/chat.uit.no`*, UiT's safe version of ChatGPT c.f. `https://en.uit.no/om/kunstigintelligens#collapse_829866`

[6] `https://github.com/flammie/omorfi/releases/tag/v0.9.11`

[7] `https://docs.google.com/`

User

I'm studying spell-checking in Finnish, can you download Alice's adventures in wonderland in Finnish from project gutenberg and produce a list of misspelt words in the text?

ChatGPT

I cannot directly download or process texts from external websites like Project Gutenberg. However, you can download the Finnish version of Alice's Adventures in Wonderland (if available) from Project Gutenberg yourself. Once you have the text, you can share it with me, and I can help analyze it for misspelled words.

User

Ok, I have attached the file here: liisa-ihmemaassa.text OTHER file

ChatGPT

Thank you for sharing the text. I can now analyze it for misspelled words. However, I need to clarify: do you want me to identify words that are not valid Finnish words, or are you looking for words that might be contextually incorrect?

User

Only words that are not valid Finnish at all, regardless of context

Figure 1: ChatGPT prompt for spell-checking and correction