Smotrom tvoja på ander drogoj verden!
Resurrecting Dead Pidgin with Generative Models:
Russenorsk Case Study

Alexey Tikhonov
Independent researcher
Berlin, Germany
altsoph@gmail.com

Sergei Shteiner
Independent researcher
Berlin, Germany
sergei.shteiner@gmail.com

Anna Bykova
LEYA, Higher School of Economics
St. Petersburg, Russia
annakholmovaia@gmail.com

Ivan P. Yamshchikov
CAIRO, THWS
Würzburg, Germany
ivan.yamshchikov@thws.de

Abstract

Russenorsk, a pidgin language historically used in trade interactions between Russian and Norwegian speakers, represents a unique linguistic phenomenon. In this paper, we attempt to analyze its lexicon using modern large language models (LLMs)[1], based on surviving literary sources. We construct a structured dictionary of the language, grouped by synonyms and word origins. Subsequently, we use this dictionary to formulate hypotheses about the core principles of word formation and grammatical structure in Russenorsk and show which hypotheses generated by large language models correspond to the hypotheses previously proposed ones in the academic literature. We also develop a "reconstruction" translation agent that generates hypothetical Russenorsk renderings of contemporary Russian and Norwegian texts.

1 Introduction

This paper is part of a relatively recent trend to use Large Language Models (LLMs) for scientific discovery in various research areas, demonstrating that these natural language processing models have unprecedented generalization capacity. To our knowledge, this is the first result that demonstrates how large language models could be effectively used in a study of dead contact languages. We first provide a brief overview of LLM-driven scientific discovery and then dive deeper into Russenorsk and its properties.

1.1 LLMs in Scientific Discovery

Natural language processing methods could inform various research hypotheses in a very broad collection of research fields, be it biochemistry Ferruz et al. (2022), human psychology Safdari et al. (2023); Sorokovikova et al. (2024), or even ancient history Yamshchikov et al. (2022); Riemenschneider and Frank (2023); Gorovaia et al. (2024); Schmidt et al. (2024). Yet this trend is boosted with the arrival of the automated Scientific Discovery Agents Jansen et al. (2024) — AI-driven systems designed to autonomously conduct scientific research by generating hypotheses, designing and executing experiments, analyzing data, and formulating conclusions without human intervention. They are used in various fields, including mathematics Lightman et al. (2023); Dong et al. (2023), social sciences Yang et al. (2023), astronomy Nguyen et al. (2023); Ciucă et al. (2023), and chemistry Jegorovè et al. (2023). Notably, the 2024 Nobel Prize in Chemistry was awarded to David Baker, Demis Hassabis, and John Jumper for their groundbreaking work in computational protein design and protein structure prediction, highlighting the transformative impact of AI-driven methods in scientific discovery Jumper et al. (2021). Large Language Models (LLMs) are increasingly being used to support scientific research, particularly in data anal-

---

[1] We used anthropic/claude-3-5-sonnet-20241022 and openai/o1-2024-12-17

ysis and hypothesis generation. In our case, we explore how LLM-based Automated Scientific Discovery Agents can assist in studying Russenorsk and gaining new insights into this historical contact language.

1.2 Russenorsk

Russenorsk is a highly unusual pidgin. de la Fuente (2020) states that its two primary lexifiers, Russian and Norwegian, contributed to its lexicon in roughly equal measure. This makes Russenorsk stand out from other pidgins. This statement is echoed in Stern (2020). Boldižarová (2021) concludes that "to understand the Russenorsk pidgin, some knowledge of the second lexifier is needed" and that "the Norwegian language skills are very beneficial but insufficient". At the same time one has to admit that we have very limited amount of sources documenting the pidgin, and depending on the choice of the analyzed documents, different researchers come to drastically different conclusions about the properties of Russenorsk. For example, Kortlandt-Leiden (2000) follows Broch and Jahr (1981) and concludes that "Russenorsk is a variant of Norwegian with an admixture of Russian foreigners' talk and elements from the native language of the speaker".

1.3 Key Contributions

This paper attempts to leverage large language models to analyze the structure of synonym formation in Russenorsk, as well as to understand the fundamental principles of phonetic and morphological transformations of words from Russian and Norwegian when incorporated into the Russenorsk lexicon.

We present three key contributions aimed at advancing the study of Russenorsk and its broader linguistic implications:

- Expanded Russenorsk Vocabulary; we introduce a newly compiled set of lexemes that extends the existing Russenorsk wordlist, offering both contemporary scholars and digital tools a more comprehensive lexical resource. This vocabulary captures variant forms and rare lexical items noted in historical sources and field observations.

- Linguistic Discovery Assistant; building on this enriched lexicon, we propose a pipeline that automates critical steps in scientific inquiry—from dictionary-based feature extrac-

tion to the generation of hypotheses and preliminary translations. This framework enables researchers to rapidly test linguistic assumptions, compare usage across contexts, and refine their understanding of Russenorsk's structural properties.

- Translation Reconstruction for Extinct Languages; finally, this paper is a case study presenting how the above pipeline can support the reconstruction of translations in a contact language no longer spoken. By leveraging parallel text alignment and informed lexical matching, our method offers a systematic approach to retrieving and interpreting content in Russenorsk, thus shedding light on its historical usage and guiding similar efforts for other extinct or endangered languages.

2 New Russenorsk Vocabulary

We present a new, extensive Russenorsk Vocabulary based on a variety of sources available in digital form.

To compile a Russenorsk vocabulary, we began by extracting lemmas from two primary sources: the online repository at https://en.wiktionary.org/wiki/Category:Russenorsk_lemmas and digitized version of Broch and Jahr (1984). This process yielded a CSV file with 691 entries. Each entry encompasses:

1. *Russenorsk form* (unique words or fixed expressions, potentially with variant spellings),

2. *Cyrillic representation* (if identified),

3. *Translations in English/Norwegian/Russian*,

4. *Part-of-speech*,

5. *Comment* (additional notes or observations),

6. *Russenorsk example* from the source text,

7. *Norwegian translation* of the same example,

8. *Russian translation* of the Norwegian example (manually refined from automatic translations),

9. *Book page reference* (if an example was found),

10. *Link* to relevant Wiktionary information,

11. *IPA pronunciation* (if available).

```
{
  "entity": {
    "english": "good morning",
    "russian": "доброе утро",
    "norwegian": "god morgen"
  },
  "classes": [],
  "pos": "phrase",
  "notes": [
    "Appears to be a blend of Russian dobroe utro
        and Norwegian god morgen",
    "Combines Russian dobra (good)
        with Norwegian morgen (morning)"
  ],
  "variants": [
    {
      "word": "dobra morgen",
      "origin": [
        {
          "lang": "Russian",
          "word": "доброе"
        },
        {
          "lang": "Norwegian",
          "word": "morgen"
        }
      ]
    }
  ]
}
```

Figure 1: An example of a resulting vocabulary entry.

After assembling the initial dictionary, we employed a pipeline (based on a Sonnet-driven system) to cluster entries with similar morphology, and subsequently grouped these by semantic proximity. This reformatting required manual review over several days to remove erroneous groupings and "hallucinated" entries. Finally, the cleaned dictionary was reviewed by additional collaborators, who discovered a few remaining mistakes that were subsequently corrected by hand.

Figure 1 provides an example of a vocabulary entry. Full vocabulary is available online[2].

3 Linguistic Discovery Assistant

Using the constructed vocabulary as input we developed a Linguistic Discovery Assistant. We propose a pipeline for hypothesis generation that leverages a large language model (LLM) as an exploratory tool for analyzing Russenorsk linguistic phenomena.

3.1 Agentic Pipeline

Let us provide an overview of the core agentic steps here. We publish the prompts, the answers, the generate hypotheses and further material regarding to facilitate further work and reproducibility[3].

(1) Lexicon and Origin Hypotheses. We first supply the LLM with our compiled Russenorsk dictionary (Section 2), prompting it to propose theories regarding the source language of each borrowed term. These requests explicitly ask the model to identify whether the etymon stems from Russian, Norwegian, or a third language.

(2) Phonetics and Morphology. Next, we reuse the same dictionary to elicit hypotheses about how the morphological and phonetic characteristics of the borrowed terms might evolve when integrated into Russenorsk. The model is thus encouraged to consider potential pathways of phonetic simplification, morphological reduction, or other structural transformations characteristic of contact languages.

(3) Grammar and Syntax. Finally, we provide the LLM with our dictionary entries alongside example sentences (the same sources used for building the dictionary; see Section 2). We prompt the model to infer possible grammatical rules, such as word-order constraints, morphological markers, or the role of function words within Russenorsk.

3.2 Ablations and Verification

To ensure that the system's outputs are neither memorized content from model pre-training nor purely "hallucinated" responses, we introduce two ablation scenarios:

1. **Empty Dictionary (Russenorsk Context).** We inform the model that we are studying Russenorsk but provide an empty dictionary. Any generated hypotheses likely stem from pre-trained data.

2. **Fictitious Pidgin.** We inform the model that the target language is a non-existent pidgin and again supply no dictionary. Here, any hypotheses should reflect purely inferential or generic reasoning rather than grounded linguistic knowledge.

By comparing the model's proposals in these ablation settings with those generated under full-information conditions, we can isolate the role of the dictionary in shaping its hypotheses. Lastly, we juxtapose the system's final output with documented linguistic insights from over a century of Russenorsk research. Although historical sources are occasionally inconsistent, mapping the model's hypotheses to these scholarly discussions highlights both convergences and points of contention, offering a systematic approach to modern Russenorsk inquiry.

4 Properties of Russenorsk

In this section, we present key syntactic, morphological, phonetic, and lexical features that characterize Russenorsk, drawing on prior research Perekhval'skaya (1987); Davydov et al. (1986); Broch (1927); de la Fuente (2020). Most of the

---

[2]https://github.com/altsoph/RusNor/blob/main/vocabulary/rusnor.v0.1.json

[3]https://github.com/altsoph/RusNor/

human-proposed properties are also proposed by the discovery agent. We publish additional agentic hypotheses, yet further verification of those is beyond the scope of the current study[4].

4.1 Grammar and Syntax

Word Order. Although Russenorsk exhibits a relatively free word order, it appears to favor a Subject–Object–Verb (SOV) pattern Perekhval'skaya (1987). This flexibility is largely influenced by Russian, where free-word order is common Davydov et al. (1986).

Minimal Inflectional Morphology. Russenorsk generally lacks derivational and inflectional morphology, a characteristic feature of contact languages Perekhval'skaya (1987). Nouns do not mark case, numberor gender, and verbs lack categories such as person, number, tense, aspect, and voice Perekhval'skaya (1987).

Frozen Forms. Many Russenorsk words correspond to "frozen" morphological forms derived from Russian or Norwegian. These forms are adopted as-is without further inflection Broch (1927).

Verb Suffix -om. One of the most distinctive features of Russenorsk is the frequent -om suffix on verbs, which serves to mark verbal categories Perekhval'skaya (1987); de la Fuente (2020). For many researchers, this morphological pattern represents Russenorsk's sole systematic (suffix-based) grammatical form Broch (1927).

Absence of Articles. Russenorsk does not exhibit an article system, closely mirroring Russian's article-free structure Broch (1927).

Universal Preposition [på]. A single multifunctional preposition — commonly realized as [på], [paa], or [pa] — fulfills a wide range of grammatical and semantic roles (e.g., locative, temporal, case-marking) Perekhval'skaya (1987); Broch (1927). Its etymology combines both Norwegian [på] and Russian [po], reflecting the language's mixed origins.

Interrogatives and Other Function Words. Russenorsk employs several well-attested strategies for forming questions (Broch, 1927): (i) an initial Wh-word (*kak*, *skoro*, etc.); (ii) a declarative

clause interpreted as a question solely by intonation (e.g. *tvoja fisk kupom?*); (iii) inversion of verb and subject (*vil ju på moja stova ... ?*); and (iv) insertion of the Russian interrogative particle *li* .

The particle *li*, borrowed from Russian, can stand alone after the focused phrase or appear fused inside lexicalized forms such as *mangeli / nogli* 'how many?', reflecting its origin in the frequent Russian pattern *mnogo-li* (Broch, 1927). Such compounds show that speakers had re-analysed *li* as a productive—if limited—morpheme.

Russenorsk also borrowed the Norwegian discourse particle *kanske* 'maybe'. Placed sentence-initially, *kanske* functions as a politeness hedge or tentative marker (e.g. *kanske tvoja vil glas tšai?* 'perhaps you'd like a glass of tea') and is often semantically bleached, bordering on an "empty" pragmatic filler (Broch, 1927).

Negation. Russenorsk expresses negation using forms derived from Russian [njet] and Norwegian [ikke], often interchangeably. In some cases, [njet] may appear as a simple negative marker as well as an indicator of "no" Broch (1927).

Lack of Copula. Russenorsk generally does not employ a copular verb akin to "to be." This absence is another hallmark of its simplified morphological structure Perekhval'skaya (1987).

4.2 Phonetics and Morphology

Depalatalization. As noted in de la Fuente (2020), Russenorsk, like many other pidgins and creoles, exhibits a strong tendency toward the neutralization of palatal elements. This is exemplified by the loss of the palatal element in the transformation of the Russian -jom into Russenorsk -om.

Convergence of Russian and Norwegian Grammatical Endings. According to de la Fuente (2020), the noun marker -*a* is a fusion of elements from Russian and Norwegian, the verb marker -*om* combines the Russian first-person plural and hortative -*jom*, possibly influenced by Cape Dutch pidgin, and the preposition *po* is a blend of the Russian *po* and the Norwegian *på*.

Tendency to add a final syllable. As noted in Hirnsperger (2012), Russenorsk exhibited a tendency to append an additional syllable to words, despite the fact that monosyllabic words are common in both Russian and Norwegian. For example, the Norwegian word *fisk* ('fish') became *fiska* in Russenorsk.

---

[4] https://github.com/altsoph/RusNor/tree/main/responses/

Elimination of front rounded vowels. Belikov (2003) provides the following examples, illustrating this tendency, which leads to phonetic simplification. For instance, Norwegian *tønde* ('barrel') became Russenorsk *tunna*, and Norwegian *søndag* ('Sunday') transformed into Russenorsk *sondag*.

Devoicing of final voiced consonants. Belikov (2003) provides the following example, illustrating this tendency: Norwegian *hav* ('sea') became Russenorsk *gaf*.

Possible Russian-to-Norwegian phonetic accommodation ([x] → [k]). Belikov (2003) states:

> "It is highly probable that the Russians also simplified the phonetics of words of Russian origin, making their pronunciation closer to Norwegian standards (changing, in particular, x to k)."

Replacement of glottal [h] by [g]. Belikov (2003) provides examples of this phenomenon, such as the transformation of Norwegian *halv* ('half') into Russenorsk *hal* or *gall*, and Norwegian *hav* ('sea') into Russenorsk *gav*.

Splitting up certain Russian consonant clusters. Hirnsperger (2012) notes that certain Russian consonant clusters were split to better align with Norwegian phonetic patterns. For example, Russian *mnogo li* ('many?') was adapted in Russenorsk as *nogoli*.

4.3 Lexicon and Origin Hypotheses

Fish and Maritime Terms from Norwegian. Words related to fishing and maritime trade were predominantly borrowed from Norwegian Davydov et al. (1986) Davydov et al. (1986).

Goods of Inland Origin from Russian. Terms for goods of inland origin, such as flour, bread, and fabric, often have Russian roots Davydov et al. (1986).

Greetings and Farewells, Possibly More Frequent from Russian. Hayden (2019) notes that all recorded greetings and farewells in Russenorsk contain a Russian element. However, their prevalence may simply reflect the Norwegian provenance of most recorded texts.

Preference for Shorter or Simpler Words. Since the morphological and phonetic structures of Russenorsk are highly simplified, it predominantly retains lexical items that are the easiest to pronounce and understand, regardless of their original linguistic source Perekhval'skaya (1987).

Synonyms of Dual Origin (Russian and Norwegian). For many concepts in Russenorsk, the existing corpus of texts records the presence of two synonyms, typically one of Russian origin and the other of Norwegian origin Perekhval'skaya (1987); Hayden (2019); Broch (1927).

Single-Origin Synonyms. Hayden (2019) states:

> "Single-source synonym appears to have gone unnoticed in the literature, though synonym groups of this type also occur, albeit not as frequently as multi-source synonym groups."

Merging of Similar Words from Both Source Languages. When the source languages contain words that are phonetically similar and semantically appropriate, they sometimes merge into a single Russenorsk term—a process referred to as "dual etymology" Hayden (2019).

A Small Percentage of Words from Other Languages. Russenorsk contains a small number of words that likely originate from Dutch, Low German, High German, and English Hayden (2019).

Self-Glossing. Sometimes, synonyms—either from both source languages or even from the same language—are used together to increase the likelihood of correctly conveying the message. Examples include *korosjo dobra* ('good good'), *stari gammel* ('old old'), *lita gran nemnožko* ('a little a little'), and *nokka lite* ('a little a little') Hayden (2019).

Founder Effect. According to Hayden (2019), certain words may establish themselves as the sole widely used term, possibly due to a founder effect or because one group used them more frequently when initiating contact.

5 Results of Agentic Discovery

The agent presented in Section 3 produced the following key findings:

- **Grammar and Syntax** Overall, nearly all hypotheses align with the observed features of Russenorsk. The only contradiction is that the models suggest that Russenorsk exhibits a preference for SVO word order, whereas Perekhval'skaya (1987) proposes an SOV

structure as possibly slightly more characteristic.

A quick qualitative scan of our corpus suggests that both SVO and SOV clauses occur with comparable frequency, with SVO only slightly more common overall. The persistence of SOV remains noteworthy, however, because that order is not characteristic of either Norwegian or Russian.

- **Phonetics and Morphology** Overall, all hypotheses of the models align with the features outlined above. The only phonetic characteristic not identified by the models is the replacement of glottal *[h]* by *[g]*.

- **Lexicon and Origin Hypotheses** Most lexical hypotheses produced by the models align with the features outlined above. However, the models overlook several attested patterns: (i) the tendency for greetings and farewells to originate primarily from Russian, (ii) the presence of single-origin synonyms, and (iii) the self-glossing strategy in which near-synonymous words are doubled to enhance comprehension. Additionally, two hypotheses—*Fish names from the language of the primary catch zone* and *Preference for words ending in vowels*—lack direct confirmation in the existing literature but appear plausible.

Table 1 allows to see which properties of Russenorsk mentioned in the literature were suggested by the generative models prompted along the agentic pipeline described in Section 3. It also shows which of these hypotheses were generated by the agent without any dictionary information, yet with information that we are interested in Russenorsk, as well as by the agent that had to dictionary and no information on which specific pidgin we want to study.

Claude/Sonnet achieves marginally higher recall than o1. At the same time, the union of their outputs provides broader coverage than either model individually, underscoring a useful cumulative effect when both systems are consulted in parallel.

6 "Reconstruction" Translation for Extinct Languages

The final component of our study demonstrates how the pipeline introduced in Section 3 can be extended to "reconstruct" translations for a contact language no longer in active use. Russenorsk serves here as a case study, illustrating the potential for systematic alignment of parallel text, lexical matching, and informed linguistic hypotheses to recover and interpret historical content. Although this approach is not strictly falsifiable, it offers a practical methodology for scholars working with extinct or endangered languages. The prompts for such a "reconstruction" translation are published online[5].

6.1 Pipeline Overview

We integrate each stage of the linguistic discovery workflow into a translation task:

1. **Dictionary Integration.** We supply the model with the Russenorsk lexicon (Section 2) to anchor the translation in the existing word inventory.

2. **Examples and Hypotheses.** We present the example sentences[6] and the hypotheses generated in Section 3, covering source-language identification, morphological/phonetic adaptation, and grammatical structure.

3. **Source.** We provide segments of original Russenorsk text that require translation into modern Russian or Norwegian.

4. **Prompted Translation.** The model is instructed to reference all prior information (dictionary entries, morphological/phonetic rules, and grammatical constraints) to produce a Russenorsk translation—preferring known lexemes over inventing new forms.

6.2 Instruction

The pipeline described in the previous subsection results in a "reconstruction" translation into hypothetical historical Russenorsk. While the exact fidelity of the final output is challenging to verify (due to the language's extinct status and limited corpus), the method provides a principled approach that combines:

- **Parallel Alignment:** Aligning segments from Russenorsk source text with corresponding pieces in modern languages.

---

[5]https://github.com/altsoph/RusNor/blob/main/tr_exp/tr_test1.rusnor.prompt.tpl
[6]https://github.com/altsoph/RusNor/blob/main/vocabulary/sentence_examples.txt

Hypothesis	Sonnet	o1	Empty Dictionary	Fictitious Pidgin
Grammar and Syntax				
Word Order	✗	✗	✗	✓
Minimal Inflectional Morphology	✓	✓	✓	✗
Frozen Forms	✓	✓	✗	✗
Verb Suffix -om	✓	✓	✗	✗
Absence of Articles	✓	✓	✗	✓
Universal Preposition [på]	✓	✗	✗	✗
Interrogatives and Other Function Words	✓	✓	✗	✗
Negation	✓	✓	✗	✗
Lack of Copula	✓	✓	✗	✗
Phonetics and Morphology				
Depalatalization	✓	✓	✗	✗
Convergence of Grammatical Endings	✓	✓	✗	✗
Tendency to Add a Final Syllable	✓	✓	✗	✗
Elimination of Front Rounded Vowels	✓	✓	✗	✗
Devoicing of Final Voiced Consonants	✓	✗	✗	✗
Russian-to-Norwegian Phonetic Accommodation	✓	✗	✗	✗
Replacement of Glottal [h] by [g]	✗	✗	✗	✗
Splitting up Russian Consonant Clusters	✓	✓	✓	✓
Lexicon and Origin Hypotheses				
Fish and Maritime Terms from Norwegian	✓	✓	✗	✗
Goods of Inland Origin from Russian	✓	✓	✗	✗
Greetings and Farewells from Russian	✗	✗	✗	✗
Preference for Shorter or Simpler Words	✓	✓	✗	✗
Synonyms of Dual Origin	✓	✗	✗	✓
Single-Origin Synonyms	✗	✗	✗	✗
Merging of Similar Words from Both Languages	✓	✗	✗	✗
A Small Percentage of Words from Other Languages	✗	✓	✗	✗
Self-Glossing	✗	✗	✗	✗
Founder Effect	✓	✓	✗	✓

Table 1: Overview of agentic discovery hypotheses across ablations.

- **Lexical Matching:** Prioritizing existing dictionary forms to maintain consistency and reduce unnecessary neologisms.

- **Hypothesis-Driven Variation:** Incorporating morphological and phonetic adaptation rules derived from Section 3.

Here is an example of an obtained "reconstruction" Russenorsk translation for a traditional Pomor tale. For the English version we refer the reader to the Appendix A.

Stannom starik på staraja kona. Odin ras starik på skoga spaserom på trovva

rubom. Gak on vil trea rubom, trea prosom: "Njet rubom moja, saa moja på tvoja ønska dast." Starik prosom rik bli. På stova kom - datsja nova stannom, grot kleba stannom, på diengi bolsjoi. Men staraja kona på klokka sprækom: "Kak moja rik levom, men folka på moja njet ærom!" Saa kona vil starik på makta stannom. Trea dast etta ønska. Men kona njet dobra - på makta stannom drogoj makta. Kona prosom starik på trea sprækom: først på bolsjoi makta, saa på større makta, saa på grot makta, saa på tsar, saa på boga stannom. Trea dast altsamma, men på siste ønska starik på kona på bjørna stannom.

By synthesizing these elements, our pipeline provides a coherent mechanism for generating Russenorsk translations that reflect plausible lexical and grammatical patterns. While certain inferences remain speculative, the overarching framework illuminates both the linguistic "gaps" in historical texts and the utility of modern LLM-driven tools in bridging them.

6.2.1 A cautious, metric-based evaluation of reconstructed translation quality

To obtain a first, *strictly indicative* estimate of translation quality, we extracted 35 well-formed sentences for which complete triplets were available in Russian (*ru*), Norwegian (*no*) and legacy Russenorsk (*rn*).[7] All automatic translations were produced with `anthropic/claude-3-5-sonnet-20241022`; we used the rule hypotheses (introduced in Section 3), derived with the same model. For every sentence we asked the model to produce four directions: *ru→rn*, *no→rn*, *rn→no* and *rn→ru*.

Prompt leakage and ablations Because the prompt relies on artefacts—the lexicon, the example sentences, and the hypotheses—derived largely from the same historical sources as our 35-sentence benchmark, the model gains a potential advantage, further amplified by fragments it may have memorised during pre-training. To gauge the scale of this *prompt leakage*, we run a three-step ablation study, successively removing (1) the examples, (2) the lexicon, and (3) the rules while keeping every other setting unchanged.

---

[7]The set is released at https://github.com/altsoph/RusNor/blob/main/tr_exp/tr_bench.json.

Automatic metric Conventional word-level metrics such as BLEU (Papineni et al., 2002) are unreliable here because both Russenorsk and Russian exhibit heavy inflection, optional diacritics, and tokenisation ambiguity. We therefore restrict ourselves to the character-level F-score chrF (Popović, 2015), which correlates well with human judgements under such conditions. Before scoring, both references and hypotheses are lower-cased and stripped of diacritics. All chrF values are computed with the `sacrebleu` toolkit (Post, 2018) to ensure fully reproducible numbers.

Reference points As a lower bound we compute chrF between the original Russenorsk sentences and their Norwegian counterparts (as they share the alphabet and some part of the lexicon); the score is ≈ 17, which we treat as "chance-level comprehension."

For a practical upper bound we use the strongest public OPUS-MT models trained on modern data. The official benchmark reports **chrF 0.418** for Russian→Norwegian (opus-mt-ru-no) (Helsinki-NLP.a) and **chrF 0.400** for Norwegian→Russian (opus-mt-no-ru) (HelsinkiNLP.b). These figures therefore set an optimistic ceiling for automatic translation on this language pair; our Russenorsk reconstructions remain well below that threshold.

Interpretation We now interpret the chrF scores summarised in Table 2.

- The dramatic gap between directions *into* Russenorsk and *out of* Russenorsk under the full prompt indicates that the model often copies or lightly paraphrases example sentences when they are available.

- Removing those examples lowers *ru→rn* and *no→rn* by roughly 30–40 chrF points, yet they still outperform the 17-chrF baseline, showing that the dictionary and rules convey additional, non-trivial signal.

- With only high-level rules left, performance drops to ≈ 22–23 chrF for *ru/no→rn* but remains slightly above baseline, hinting at a limited ability to *compose* unseen Russenorsk forms.

- In the zero-resource setting (*none*) the model still reaches 33–40 chrF for *rn→ru/no*, presumably because many tokens are transparent cognates, and the model is well familiar with Norwegian and Russian. Translation

18163

Table 2: chrF scores (↑) for the four translation directions under incremental ablations. The first row shows a lower-bound reference obtained by comparing the original Russenorsk sentences with their Norwegian counterparts. Red results demonstrate the leakage of exemple sentences. Purple results show the model is well familiar with Norwegian and Russian, and many tokens are transparent cognates.

Prompt contents	ru→rn	no→rn	rn→ru	rn→no
baseline (rn vs. no)	–	17.8	–	17.4
dictionary + examples + rules	67.7	56.8	37.5	46.2
– examples	**29.2**	**32.0**	38.5	45.8
rules only	22.4	22.7	33.6	38.6
none (memory test)	19.7	20.4	33.2	40.0

into Russenorsk, however, collapses to near-baseline values, confirming that productive generation requires explicit lexical support.

Overall, the experiment shows incremental, measurable gains from each resource layer while underscoring the ease with which large language models can over-fit to illustrative material. Given the tiny sample and the exploratory nature of chrF in a pidgin context, these figures should be considered a *sanity check* rather than a definitive quality assessment. A proper human evaluation — ideally with experts in Scandinavian contact languages — might be an essential next step.

6.3 Outlook for Extinct and Endangered Languages

Although demonstrated here with Russenorsk, this translation-reconstruction approach can extend to other contact languages or endangered linguistic varieties with similarly sparse documentation. By systematically integrating dictionary resources, morphosyntactic hypotheses, and historical text samples, researchers can better preserve linguistic heritage and potentially uncover previously overlooked facets of these languages' structures.

7 Conclusion

This paper demonstrates how modern large language models could be used in linguistics of low-resource dead languages and pidgins. We create the first structured dictionary of Russenorsk and make it available under an open-source license. The lexicon primarily consists of borrowed terms from Russian and Norwegian, with occasional influences from other languages. The distribution of terms reflects the functional nature of the pidgin as a trade language. We also publish a linguistic discovery agent and demonstrate its capability to generate viable linguistic hypotheses about Russenorsk. We believe similar ideas could be applied to other dead

or low-resource pidgin languages. Finally, we provide the first automated pipeline for "reconstructive" translation into Russenorsk.

Russenorsk æ god språk for folk. Vi snakke russisk, norsk, og litt mer, for handel og vennskap. Det viser hvordan folk kan lage nye språk sammen, lett og rask.

Limitations

Given the limited documentation of Russenorsk, the output of an agentic discovery pipeline is highly speculative and needs further investigations that include corpus-based analyses, comparative studies with other pidgins, and investigations into the sociolinguistic contexts of its use. The same limitations restricts the "reconstructed" translation pipeline. We use "reconstructed" in the quotation marks to draw attention of the reader to the fact that the resulting output could hardly be called a translation in the common sense of the word.

Ethics Statement

This paper complies with the ACL Ethics Policy. The manual annotation of the obtained results was carried out by the co-authors of the paper, thus we acknowledge the potential for annotator bias, which could impact dataset consistency.

References

Vladimir Belikov. 2003. Some fragments of russenorsk grammar.

Katarína Boldižarová. 2021. The contemporary norwegians' understanding of russenorsk.

Ingvild Broch and Ernst Håkon Jahr. 1984. *Russenorsk – et pidginspråk i Norge [Russenorsk – a Pidgin Language in Norway]*, 2nd edition. Novus, Oslo.

Ingvild Broch and Ernst Hakon Jahr. 1981. Russenorsk: Et pidginspråk i norge. *Tromsø-Studier i Språkvitensk*, 3. In Norwegian.

Olaf Broch. 1927. Russenorsk. *Archiv für Slavische Philologie*, 41:81–130.

Ioana Ciucă, Yuan-Sen Ting, Sandor Kruk, and Kartheik Iyer. 2023. Harnessing the power of adversarial prompting and large language models for robust hypothesis generation in astronomy. *Preprint*, arXiv:2306.11648.

A. N. Davydov, V. N. Ponomarenko, and A. A. Kuratova. 1986. Russenorsk - arkticheskiy pidzhin evropy. In Russian.

José Andrés Alonso de la Fuente. 2020. On russenorsk -om in particular and on etymology and creolistics in general.

Qingxiu Dong, Li Dong, Ke Xu, Guangyan Zhou, Yaru Hao, Zhifang Sui, and Furu Wei. 2023. Large language model for science: A study on p vs. np. *Preprint*, arXiv:2309.05689.

Noelia Ferruz, Steffen Schmidt, and Birte Höcker. 2022. Protgpt2 is a deep unsupervised language model for protein design. *Nature communications*, 13(1):4348.

Svetlana Gorovaia, Gleb Schmidt, and Ivan P Yamshchikov. 2024. Sui generis: Large language models for authorship attribution and verification in latin. In *Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities*, pages 398–412.

Bryan Hayden. 2019. Was russenorsk a continuum?

HelsinkiNLP.a. opus-mt-ru-no model card. https://huggingface.co/Helsinki-NLP/opus-mt-ru-no. ChrF2 score = 0.418; accessed 19 May 2025.

HelsinkiNLP.b. opus-mt-no-ru model card. https://huggingface.co/Helsinki-NLP/opus-mt-no-ru. ChrF2 score = 0.400; accessed 19 May 2025.

Markus Hirnsperger. 2012. „pidgin - russisch" - am beispiel von „russenorsk". Arbeitsgemeinschaft Arktis und Subarktis (A.A.S.).

Peter Jansen, Marc-Alexandre Côté, Tushar Khot, Erin Bransom, Bhavana Dalvi Mishra, Bodhisattwa Prasad Majumder, Oyvind Tafjord, and Peter Clark. 2024. DISCOVERYWORLD: A virtual environment for developing and evaluating automated scientific discovery agents. *Preprint*, arXiv:2406.06769. Accepted to NeurIPS 2024 (Benchmark Track, Spotlight).

Aistė Jegorovė, Jianxing Xia, Matas Steponaitis, Maryte Daskeviciene, Vygintas Jankauskas, Alytis Gruodis, Egidijus Kamarauskas, Tadas Malinauskas, Kasparas Rakstys, Khalid A. Alamry, Vytautas Getautis, and Mohammad Khaja Nazeeruddin. 2023. Branched fluorenylidene derivatives with low ionization potentials as hole-transporting materials for perovskite solar cells. *Chemistry of Materials*, 35(15):5914–5923.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. 2021. Highly accurate protein structure prediction with alphafold. *Nature*, 596:583–589.

Frederik Kortlandt-Leiden. 2000. On russenorsk. *Amsterdamer Beiträge zur älteren Germanistik*, 54:123.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Improving mathematical reasoning with process supervision.

Tuan Dung Nguyen, Yuan-Sen Ting, Ioana Ciucă, Charlie O'Neill, Ze-Chang Sun, Maja Jabłońska, Sandor Kruk, Ernest Perkowski, Jack Miller, Jason Li, Josh Peek, Kartheik Iyer, Tomasz Różański, Pranav Khetarpal, Sharaf Zaman, David Brodrick, Sergio J. Rodríguez Méndez, Thang Bui, Alyssa Goodman, Alberto Accomazzi, Jill Naiman, and Jesse Cranney. 2023. AstroLLaMA: Towards specialized foundation models in astronomy. *Preprint*, arXiv:2309.06126.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the ACL*, pages 311–318.

Elena V. Perekhval'skaya. 1987. Russenorsk kak primer pervonachal'nogo e'tapa formirovaniya pidzhina. In Russian.

Maja Popović. 2015. chrf: character n-gram f-score for automatic machine translation evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395.

Matt Post. 2018. A Call for Clarity in Reporting BLEU Scores. In *Proceedings of the Third Conference on Machine Translation (WMT)*, pages 186–191.

Frederick Riemenschneider and Anette Frank. 2023. Exploring large language models for classical philology. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15181–15199.

Mustafa Safdari, Greg Serapio-García, Clément Crepy, Stephen Fitz, Peter Romero, Luning Sun, Marwa Abdulhai, Aleksandra Faust, and Maja Matarić. 2023. Personality traits in large language models. *arXiv preprint arXiv:2307.00184*.

Gleb Schmidt, Veronica Vybornaya, and Ivan P. Yamshchikov. 2024. Fine-tuning pre-trained language models for authorship attribution of the pseudo-dionysian ars rhetorica. In *Proceedings of the Computational Humanities Research Conference 2024, Aarhus, Denmark, December 4-6, 2024*, volume 3834 of *CEUR Workshop Proceedings*, pages 369–385. CEUR-WS.org.

Aleksandra Sorokovikova, Sharwin Rezagholi, Natalia Fedorova, and Ivan Yamshchikov. 2024. Llms simulate big5 personality traits: Further evidence. In *Proceedings of the 1st Workshop on Personalization of Generative AI Systems (PERSONALIZE 2024)*, pages 83–87.

Dieter Stern. 2020. Russian pidgin languages. *Encyclopedia of Slavic languages and linguistics online*.

Ivan Yamshchikov, Alexey Tikhonov, Yorgos Pantis, Charlotte Schubert, and Jürgen Jost. 2022. Bert in plutarch's shadows. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6071–6080.

Zonglin Yang, Xinya Du, Junxian Li, Jie Zheng, Soujanya Poria, and Erik Cambria. 2023. Large language models for automated open-domain scientific hypotheses discovery. *Preprint*, arXiv:2309.02726.

A Appendix: Tale in English

Traditional Pomor tale about an old man and a tree. Translation into English:

> An old man lived with an old woman. One day the old man went into the forest to chop wood. And when he wanted to cut down another tree, it asked not to be cut down, but in return - to fulfill any wish. The old man decided to ask for wealth. He comes home - the hut has become new, there are many domestic animals, bread will last for a long time, "there is more than enough money." But after a while the old woman began to say: "Even though we live richly, what is the point if people do not respect us!", and that it would be good for the grandfather to become a burgomaster. The tree fulfills this wish as well. But the old woman is again dissatisfied - after all, there is power over the burgomaster too. She suggests that the grandfather ask the tree to make him first a master, then a colonel, a general, a king, and then God himself. The tree fulfills all this, but in response to the last request, it turns the old man and the old woman into bears.