Open, Reproducible Morphology Probes for Plains Cree

Anonymous submission

Abstract

We present a minimal, fully open baseline for morphology-aware evaluation in Plains Cree (nêhiyawêwin) designed for constrained settings. The goal is a recipe that community partners and researchers can run on a single workstation with a stock PyTorch plus Transformers environment and no additional installs, while producing machine-readable artifacts for audit and reuse. We treat a compact open causal language model as a zero-shot probe and evaluate two tasks that reflect real linguistic structure: (1) reinflection given a lemma and a plus-delimited feature bundle, and (2) analysis that outputs a plus-delimited segmentation with feature tags. Tag conventions follow GiellaLT-style resources to keep outputs interpretable. Decoding is greedy. Metrics are exact-match accuracy and average Levenshtein distance for reinflection, and Jaccard overlap over tag sets for analysis.

Using a small curated gold set (reinflection n=6, analysis n=6), we find limited morphology competence. Reinflection reaches accuracy 0.17 with average edit distance 3.17. By mood, Ind outperforms Cnj (accuracy 0.33 vs 0.00; AvgED 2.00 vs 4.33). By person, only 3Sg yields any exact matches (accuracy 0.50, AvgED 1.50). Analysis averages Jaccard 0.00 overall and in both Ind and Cnj, driven by lemma copying, missing conjunct morphology, and prompt echo that replaces structured analyses with meta-text. These results establish a clear, reproducible baseline and pinpoint failure modes to target with format-constrained prompting, a few in-context exemplars, and lightweight FST-assisted checks while maintaining an open-only constraint.

Background

Context

Work on Canadian Indigenous languages in NLP has concentrated where aligned corpora and shared tasks exist. Inuktut has seen the most traction due to the Nunavut Hansard parallel corpus and its inclusion in the WMT20 news translation task. Together these established a widely used IU↔EN benchmark and surfaced issues with script handling and domain shift (Joanis et al. 2020; Barrault et al. 2020; wmt 2020; Knowles et al. 2020; Hernandez and Nguyen 2020).

Industry deployments and limits

Platform deployments increased visibility and access. Google Translate added Inuktut support, and Microsoft introduced Inuktitut in Translator and later publicized TTS voices (Caswell 2024; Microsoft News Center Canada 2021; Kirk 2024). These releases matter for users, yet they rarely ship with reusable baselines, detailed error analyses, or artifacts that are easy for community partners to audit and extend.

Multilingual MT at scale

Large multilingual efforts such as No Language Left Behind expanded coverage to 200 languages and released widely used checkpoints with public documentation of data curation and evaluation (NLLB Team et al. 2022; Costa-jussà et al. 2024). These projects optimize for sentence-level MT quality. They offer limited insight into morphological competence for polysynthetic families like Algonquian and Inuit, and robust evaluation beyond Inuktut remains sparse.

Morphology resources and community infrastructure

Rule-based and community-aligned infrastructure makes morphology evaluation feasible. For Plains Cree, the GiellaLT ecosystem provides finite-state analyzers and generators with shared engineering conventions (Moshagen et al. 2023; GiellaLT 2025). The *itwêwina* dictionary anchors lexicographic practice used by speakers and educators (ALT-Lab, University of Alberta 2021). For Ojibwe, recent FST work lowers the barrier to analysis and generation (Hammerly et al. 2025). For Inuktitut, neural-augmented analyzers illustrate coverage challenges in polysynthetic morphology (Micher 2017). Separately, ReadAlong Studio shows practical, licensed alignment workflows for Indigenous audiobooks that can support future evaluation sets (Littell et al. 2022).

Underserved-language perspective

Broader surveys highlight structural inequities in data, compute, and evaluation for low-resource and Indigenous languages (Joshi et al. 2020). Indigenous-focused modeling such as IndT5 shows promise under sparse data, but comprehensive, reproducible LLM evaluations for Canadian Indigenous languages remain uncommon (Nagoudi et al. 2021).

Gaps

- **Coverage beyond Inuktut.** Public, standardized evaluation concentrates on IU↔EN MT. Systematic assessment for Cree and Ojibwe is limited (Barrault et al. 2020).
- **Morphology-aware evaluation.** Sentence-level MT metrics do not test whether models generate or analyze the correct inflectional feature bundles used by speakers. Finite-state resources exist, yet they are rarely paired with LLMs in open, compute-constrained probes (Moshagen et al. 2023; GiellaLT 2025; Hammerly et al. 2025).
- **Open, runnable baselines.** Many demonstrations depend on closed APIs or heavy stacks. Community partners need artifacts that run locally with permissive licenses and minimal environment friction (NLLB Team et al. 2022).

Scope of this paper

We target a narrow, high-signal objective that prioritizes reproducibility. We focus on Plains Cree morphology and operate under an open-only, no-new-installs constraint. We evaluate two tasks:

1. **Reinflection** given a lemma and a plus-delimited feature bundle, scored by exact match and average edit distance.
2. **Analysis** given a surface form, producing a plus-delimited segmentation and features, scored by Jaccard overlap over tag sets.

We treat a compact open causal LM as a black-box probe. We align tag conventions with the GiellaLT ecosystem and community dictionaries to keep outputs interpretable by practitioners (Moshagen et al. 2023; GiellaLT 2025; ALTLab, University of Alberta 2021). The goal is not state of the art. The goal is a clean, runnable baseline that labs and community partners can re-run, critique, and extend.

Summary Prior work gives strong MT resources for Inuktut and mature finite-state morphology for Cree and Ojibwe. What is missing is open, reproducible, morphology-aware evaluation for Cree. We fill that by running an offline probe on reinflection and analysis with open weights and simple, auditable metrics tied to community conventions.

Methodology

Study design and constraints

- **Scope**: evaluate morphology for Plains Cree (nêhiyawêwin) only. We do not train or fine tune models.
- **Open only**: weights and data must be publicly available under permissive licenses. No closed APIs.
- **No new installs**: rely on a stock PyTorch + Transformers environment.
- **Single machine**: inference on one workstation with deterministic decoding.
- **Community alignment**: tag conventions follow GiellaLT style resources for Cree (Moshagen et al. 2023; GiellaLT 2025); lemmas are grounded in community lexicography where possible (ALTLab, University of Alberta 2021).

Task	Input	Output	Primary score
Reinflection	ℓ, b	surface $\hat{y}$	exact match, avg edit distance
Analysis	x	analysis $\hat{a}$	Jaccard over tag sets

Table 1: Tasks and scores. Plus delimited bundles follow GiellaLT conventions (GiellaLT 2025; Moshagen et al. 2023).

Tasks

We evaluate two morphology tasks that expose complementary behavior.

Reinflection Given a lemma ℓ and a plus delimited feature bundle b (e.g., `V+AI+Ind+Prs+1Sg`), generate a surface form $\hat{y}$.

$$(\ell, b) \;\mapsto\; \hat{y}$$

Analysis Given a surface form x, produce a plus delimited analysis $\hat{a}$, which may contain morphemes and tags in a single linearization.

$$x \;\mapsto\; \hat{a} = m_1 + m_2 + \cdots + \tau_1 + \cdots$$

Data

Gold items We curate compact test items for both tasks that are non sacred and reflect widely taught paradigms. Each reinflection item is a triple $(\ell, b, y^\star)$ with one gold surface form $y^\star$. Each analysis item is $(x, \mathcal{A}^\star)$ where $\mathcal{A}^\star = \{a_1, \ldots, a_K\}$ is a small set of acceptable analyses to accommodate orthographic or tagging variants used in practice.

Tagging conventions Feature bundles and segmentation use the same atom names and ordering patterns as the Cree resources in GiellaLT to make outputs legible to practitioners and compatible with future automated checks (GiellaLT 2025; Moshagen et al. 2023). Lemma spelling and example forms are cross checked against *itwêwina* (ALTLab, University of Alberta 2021).

Model and prompting

Probe model We treat a compact open causal LM as a black box probe. No gradient updates are applied. Decoding is greedy with temperature 0 and no sampling.

Prompts We use short instruction style prompts that elicit a single line answer.

- **Reinflection prompt**: *You are a morphological generator for Plains Cree. Given a lemma and a feature bundle in plus delimited tags, output only the inflected surface form. Lemma: ℓ Tags: b Form:*
- **Analysis prompt**: *You are a morphological analyzer for Plains Cree. Segment the word into morphemes using '+' and include tags. Output only one line. Word: x Analysis:*

No few shot examples are included. We strip any model preamble by taking the first line following the cue token (`Form:` or `Analysis:`).

Normalization

Before scoring we apply a minimal normalization function $\mathcal{N}$:

- Unicode NFC, trim leading and trailing whitespace.
- Preserve diacritics and Cree orthography; no lowercasing.
- Collapse internal runs of spaces to a single space. No punctuation stripping.

All metrics operate on $\mathcal{N}(\cdot)$.

Metrics

Let $\{(\ell_i, b_i, y_i^\star)\}_{i=1}^N$ be reinflection items and $\{(x_j, \mathcal{A}_j^\star)\}_{j=1}^M$ be analysis items.

Reinflection Model prediction $\hat{y}_i$ is compared to $y_i^\star$ after normalization.

$$\text{Acc} \;=\; \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\{\mathcal{N}(\hat{y}_i) = \mathcal{N}(y_i^\star)\}$$

$$\overline{\text{ED}} \;=\; \frac{1}{N}\sum_{i=1}^{N} \text{Lev}(\mathcal{N}(\hat{y}_i), \mathcal{N}(y_i^\star))$$

where $\text{Lev}(\cdot,\cdot)$ is unit cost Levenshtein distance.

Analysis Let $S(a)$ map a plus delimited string to a set of atoms by splitting on + and removing empties. We compute best overlap against any acceptable analysis:

$$\text{Jaccard}_j \;=\; \max_{a \in \mathcal{A}_j^\star} \frac{|S(\mathcal{N}(\hat{a}_j)) \cap S(\mathcal{N}(a))|}{|S(\mathcal{N}(\hat{a}_j)) \cup S(\mathcal{N}(a))|}$$

$$\overline{\text{Jaccard}} \;=\; \frac{1}{M}\sum_{j=1}^{M} \text{Jaccard}_j$$

Failure handling

- **Empty or multi line outputs**: treat as empty after normalization.
- **Prompt echo**: if the model repeats part of the prompt, we take the substring after the final cue token.
- **Out of alphabet**: we do not penalize diacritics; otherwise all Unicode code points count toward edit distance.

Reproducibility

- **Determinism**: greedy decoding, fixed random seeds, and stable tokenization. If GPU kernels require non deterministic cuBLAS paths, we record the environment override in the artifact metadata.
- **Artifacts**: the notebook writes per item predictions and JSON summaries for each task, plus a machine readable config capturing library versions, device, and decode settings.
- **Licensing**: only public or curated gold items are used. No sacred or restricted content. Tagging conventions match open resources (GiellaLT 2025; Moshagen et al. 2023).

Task	Input	Output	Primary score
Reinflection	ℓ, b	surface $\hat{y}$	exact match, avg edit distance
Analysis	x	analysis $\hat{a}$	Jaccard over tag sets

Table 2: Tasks and scores. Plus delimited bundles follow GiellaLT conventions (GiellaLT 2025; Moshagen et al. 2023).

Limitations

This probe is a black box stress test under strict constraints. It does not replace analysis with finite state tools, nor does it claim representativeness across dialects. The small curated sets should be expanded with community input and checked against analyzer outputs in future work (GiellaLT 2025; Moshagen et al. 2023).

TLDR We evaluate Cree morphology with two tasks. Inputs and outputs use plus delimited conventions aligned with GiellaLT. A compact open causal LM generates one line answers with greedy decoding. We score reinflection by exact match and average Levenshtein, and analysis by Jaccard overlap against a small set of acceptable gold analyses. Everything runs in a stock PyTorch plus Transformers environment with recorded artifacts for reproducibility.

Methodology

Study design and constraints

- **Scope**: evaluate morphology for Plains Cree (nêhiyawêwin) only. We do not train or fine tune models.
- **Open only**: weights and data must be publicly available under permissive licenses. No closed APIs.
- **No new installs**: rely on a stock PyTorch + Transformers environment.
- **Single machine**: inference on one workstation with deterministic decoding.
- **Community alignment**: tag conventions follow GiellaLT style resources for Cree (Moshagen et al. 2023; GiellaLT 2025); lemmas are grounded in community lexicography where possible (ALTLab, University of Alberta 2021).

Tasks

We evaluate two morphology tasks that expose complementary behavior.

Reinflection Given a lemma ℓ and a plus delimited feature bundle b (e.g., `V+AI+Ind+Prs+1Sg`), generate a surface form $\hat{y}$.

$$(\ell, b) \;\mapsto\; \hat{y}$$

Analysis Given a surface form x, produce a plus delimited analysis $\hat{a}$, which may contain morphemes and tags in a single linearization.

$$x \;\mapsto\; \hat{a} = m_1 + m_2 + \cdots + \tau_1 + \cdots$$

Data

Gold items We curate compact test items for both tasks that are non sacred and reflect widely taught paradigms. Each reinflection item is a triple $(\ell, b, y^\star)$ with one gold surface form $y^\star$. Each analysis item is $(x, \mathcal{A}^\star)$ where $\mathcal{A}^\star = \{a_1, \ldots, a_K\}$ is a small set of acceptable analyses to accommodate orthographic or tagging variants used in practice.

Tagging conventions Feature bundles and segmentation use the same atom names and ordering patterns as the Cree resources in GiellaLT to make outputs legible to practitioners and compatible with future automated checks (GiellaLT 2025; Moshagen et al. 2023). Lemma spelling and example forms are cross checked against *itwêwina* (ALTLab, University of Alberta 2021).

Model and prompting

Probe model We treat a compact open causal LM as a black box probe. No gradient updates are applied. Decoding is greedy with temperature 0 and no sampling.

Prompts We use short instruction style prompts that elicit a single line answer.

- **Reinflection prompt**: *You are a morphological generator for Plains Cree. Given a lemma and a feature bundle in plus delimited tags, output only the inflected surface form. Lemma: ℓ Tags: b Form:*
- **Analysis prompt**: *You are a morphological analyzer for Plains Cree. Segment the word into morphemes using '+' and include tags. Output only one line. Word: x Analysis:*

No few shot examples are included. We strip any model preamble by taking the first line following the cue token (`Form:` or `Analysis:`).

Normalization

Before scoring we apply a minimal normalization function $\mathcal{N}$:

- Unicode NFC, trim leading and trailing whitespace.
- Preserve diacritics and Cree orthography; no lowercasing.
- Collapse internal runs of spaces to a single space. No punctuation stripping.

All metrics operate on $\mathcal{N}(\cdot)$.

Metrics

Let $\{(\ell_i, b_i, y_i^\star)\}_{i=1}^N$ be reinflection items and $\{(x_j, \mathcal{A}_j^\star)\}_{j=1}^M$ be analysis items.

Reinflection Model prediction $\hat{y}_i$ is compared to $y_i^\star$ after normalization.

$$\text{Acc} \;=\; \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\{\mathcal{N}(\hat{y}_i) = \mathcal{N}(y_i^\star)\}$$

$$\overline{\text{ED}} \;=\; \frac{1}{N}\sum_{i=1}^{N} \text{Lev}(\mathcal{N}(\hat{y}_i), \mathcal{N}(y_i^\star))$$

where $\text{Lev}(\cdot,\cdot)$ is unit cost Levenshtein distance.

Analysis Let $S(a)$ map a plus delimited string to a set of atoms by splitting on + and removing empties. We compute best overlap against any acceptable analysis:

$$\text{Jaccard}_j \;=\; \max_{a \in \mathcal{A}_j^\star} \frac{|S(\mathcal{N}(\hat{a}_j)) \cap S(\mathcal{N}(a))|}{|S(\mathcal{N}(\hat{a}_j)) \cup S(\mathcal{N}(a))|}$$

$$\overline{\text{Jaccard}} \;=\; \frac{1}{M}\sum_{j=1}^{M} \text{Jaccard}_j$$

Failure handling

- **Empty or multi line outputs**: treat as empty after normalization.
- **Prompt echo**: if the model repeats part of the prompt, we take the substring after the final cue token.
- **Out of alphabet**: we do not penalize diacritics; otherwise all Unicode code points count toward edit distance.

Reproducibility

- **Determinism**: greedy decoding, fixed random seeds, and stable tokenization. If GPU kernels require non deterministic cuBLAS paths, we record the environment override in the artifact metadata.
- **Artifacts**: the notebook writes per item predictions and JSON summaries for each task, plus a machine readable config capturing library versions, device, and decode settings.
- **Licensing**: only public or curated gold items are used. No sacred or restricted content. Tagging conventions match open resources (GiellaLT 2025; Moshagen et al. 2023).

Limitations

This probe is a black box stress test under strict constraints. It does not replace analysis with finite state tools, nor does it claim representativeness across dialects. The small curated sets should be expanded with community input and checked against analyzer outputs in future work (GiellaLT 2025; Moshagen et al. 2023).

Results

Evaluation setup

We evaluate Plains Cree morphology under the constraints in Section *Methodology*: open weights, no additional installs, single workstation, greedy decoding, and two zero shot tasks. The curated test sets contain n=6 items for reinflection and n=6 items for analysis. Results are exact match accuracy and average Levenshtein distance for reinflection, and average Jaccard overlap over plus delimited tag sets for analysis. All scores are computed on normalized text.

Overall metrics

Breakdowns by mood and person

Findings

Accuracy for reinflection is 0.17 with AvgED 3.17 on this compact set. By mood, *Ind* outperforms *Cnj* (Acc 0.33 vs

Task	Score 1	Score 2
Reinflection (n=6)	Acc = 0.17	Avg ED = 3.17
Analysis (n=6)	Avg Jaccard = 0.00	

Table 3: Overall Cree morphology metrics. Acc is exact match. ED is Levenshtein distance.

Subset	n	Acc	Avg ED
Reinflection: Ind	3	0.33	2.00
Reinflection: Cnj	3	0.00	4.33
Reinflection: 1Sg	2	0.00	4.00
Reinflection: 2Sg	2	0.00	4.00
Reinflection: 3Sg	2	0.50	1.50
Analysis (Avg Jaccard)			
Analysis: Ind	3	0.00	
Analysis: Cnj	3	0.00	

Table 4: Breakdowns by mood and person.

0.00) and shows lower AvgED (2.00 vs 4.33). By person, only 3Sg reaches any exact matches (Acc 0.50, AvgED 1.50) while 1Sg and 2Sg remain at Acc 0.00 with AvgED 4.00. Analysis averages Jaccard 0.00 overall and within both *Ind* and *Cnj*, indicating that the model either copies the surface form or emits meta text rather than a plus delimited analysis.

Error patterns

- **Lemma copying** The model frequently returns the bare lemma *mîcisow* instead of realizing person prefixes *ni-*, *ki-* or conjunct morphology, which drives errors in 1Sg and 2Sg and across *Cnj*.
- **Missing conjunct markers** Gold *ê-* plus person suffixes are often absent, leading to large edit distances for *Cnj*.
- **Prompt echo in analysis** Outputs include English meta text or unsegmented copies of the input, yielding Jaccard 0.00.

Takeaways

- A compact open causal LM is not morphology competent for Cree under zero shot prompts on this minimal set. The gap between *Ind* and *Cnj* aligns with the extra morphological load in conjunct forms.
- Failure is about content selection rather than spelling. The model ignores feature bundles and prefers the lemma, which inflates edit distance and collapses analysis scores.
- Immediate next steps that keep the same constraints: add a few in context exemplars, enforce output format in prompts, and apply light post checks that normalize hyphenation and diacritics before scoring.

Reinflection reaches Acc 0.17 with AvgED 3.17; only 3Sg shows any exact matches. Analysis averages Jaccard 0.00 due to prompt echo and unsegmented copies. This baseline cleanly exposes failure modes to target with format control, minimal few shot conditioning, and simple post checks while staying fully open and reproducible.

Discussion

Reading the numbers

The probe exhibits weak morphology competence on this compact Cree set. Reinflection accuracy is 0.17 with AvgED 3.17. The gap between *Ind* and *Cnj* suggests that the model defaults to the lemma and struggles to realize conjunct morphology. Person marking shows the same pattern: only 3Sg yields any exact matches while 1Sg and 2Sg miss the ni and ki prefixes. Analysis scores average Jaccard 0.00, indicating that the model either copies the surface form or emits meta text rather than a plus delimited analysis.

Why the probe failed

- **Objective mismatch**: the model is trained for conversational text and instruction following. Morphology generation and analysis require compositional constraints that are not reinforced by general chat data.
- **Formatting sensitivity**: without examples or hard format constraints, the model often returns explanations or echoes instructions rather than the requested one line output.
- **Polysynthesis pressure**: Cree bundles several grammatical categories. Small deviations from prefix or suffix realization produce large edit distances and zero Jaccard even when the output is close in surface form.
- **Zero shot setting**: no few shot context means the model must infer both the tag semantics and the output linearization from a single instruction.

What worked and what did not

- **Worked**: a simple, auditable pipeline. Greedy decoding with minimal prompts plus deterministic text normalization gave stable measurements that are easy to replicate.
- **Did not**: analysis prompts without format constraints. Instruction leakage dominated, producing fluent English sentences instead of structured analyses.
- **Mixed**: reinflection for 3Sg. The model occasionally returned the lemma correctly where the gold surface equals the lemma, but failed to realize person and conjunct morphology elsewhere.

Improvements that keep the same constraints

We list changes that preserve open weights, no new installs, and single machine operation.

Low effort

- **Output shaping in prompts**: enforce a leading token and an explicit stop, for example `Form:␣` followed by `<eos>`, and reject any output containing spaces or punctuation for reinflection.

- **One or two exemplars**: add one in context example per task using the same plus delimited conventions. This often suppresses instruction leakage and narrows the output space.
- **Strict post filtering**: trim to the first non empty token after the cue, strip quotes, normalize hyphenation, and drop residual English letters beyond the first token for reinflection.

Moderate effort

- **Constrained decoding at the character level**: mask out non Cree alphabet characters during generation by rejecting tokens outside a whitelist. This can be implemented with a simple logits processor in Transformers.
- **FST assisted veto as a checker**: without invoking analyzers at inference time, use small regex style acceptors that reject outputs lacking ni or ki when person tags demand them, or lacking the conjunct prefix when mood is Cnj.
- **Minimal few shot templates**: one example each for $\{$Ind, Cnj$\}$ and for $\{$1Sg, 2Sg, 3Sg$\}$ covers the combinatorics without expanding the context window significantly.

Higher effort, still open

- **Synthetic augmentation from FSTs**: generate small inflection tables for a few lemmas using open analyzers and sample balanced subsets for evaluation and in context priming.
- **Lightweight adaptation**: LoRA on a small open seq2seq model using the synthetic tables. This remains feasible on a single GPU but breaks the strict zero train rule.

Threats to validity

- **Sample size**: n=6 is a smoking gun diagnostic set, not a benchmark. Scores are sensitive to item choice.
- **Orthography and diacritics**: we preserve diacritics. Different classroom or community practices could change exact match and edit distance.
- **Tag conventions**: plus delimited bundles follow one analyzer style. Alternative conventions would need mapping to compare scores.
- **Zero shot bias**: results understate what is possible with light task conditioning. That is a design choice for reproducibility rather than an upper bound on capability.

Ethics and community considerations

- **Content selection**: only non sacred, pedagogically common forms are used. Expansion should continue under community guidance.
- **Interpretability**: aligning outputs with community analyzer conventions keeps the results legible to speakers and instructors.
- **Deployment caution**: poor morphology harms user trust. These results argue for gating morphology heavy features behind community QA, even when models appear fluent in English prompts.

Reproducibility notes

- **Environment**: stock PyTorch and Transformers. Deterministic decoding. cuBLAS determinism was disabled for generation and recorded in the run config.
- **Artifacts**: per item JSONL files and summaries are written for both tasks so that independent teams can recompute every number in this section.
- **Portability**: the notebook runs without additional installs and its outputs are paths and JSON files rather than screenshots or dashboards.

Outlook

The immediate next step is to add minimal in context examples and strict output shaping, then scale to a dozen lemmas and a few more feature bundles. The same template can be replicated for Ojibwe with parallel tag conventions. Longer term, pairing this probe with FST assisted checks and small open finetunes would provide a tractable path toward community usable morphology features while keeping licensing and compute accessible.

The probe is reproducible and informative but not morphology competent: it copies the lemma, misses conjunct morphology, and fails to produce structured analyses. Tight output control and a few exemplars are the fastest fixes within the same constraints. Larger curated sets and FST assisted checks are the next leverage points for a credible community baseline.

References

2020. Shared Task: Machine Translation of News, WMT20. https://www.statmt.org/wmt20/translation-task.html. Includes IU–EN training and dev data links.

ALTLab, University of Alberta. 2021. itwêwina: Plains Cree Dictionary. https://itwewina.altlab.app/.

Barrault, L.; Biesialska, M.; Bojar, O.; Costa-jussà, M. R.; Federmann, C.; Graham, Y.; Grundkiewicz, R.; Haddow, B.; Huck, M.; Joanis, E.; Kocmi, T.; Koehn, P.; Lo, C.-k.; Ljubešić, N.; Monz, C.; Morishita, M.; Nagata, M.; Nakazawa, T.; Pal, S.; Post, M.; and Zampieri, M. 2020. Findings of the 2020 Conference on Machine Translation (WMT20). In *Proceedings of the Fifth Conference on Machine Translation*, 1–55. Online: Association for Computational Linguistics.

Caswell, I. 2024. Google Translate Learns Inuktut. https://blog.google/intl/en-ca/company-news/technology/google-translate-learns-inuktut/.

Costa-jussà, M. R.; Cross, J.; Çelebi, O.; Elbayad, M.; Heafield, K.; Ayan, N. F.; Bhosale, S.; Edunov, S.; Fan, A.; Gao, C.; the NLLB Team; et al. 2024. Scaling neural machine translation to 200 languages. *Nature*.

GiellaLT. 2025. lang-crk: Finite-state and Constraint Grammar resources for Plains Cree. https://github.com/giellalt/lang-crk.

Hammerly, C.; LeVasseur, I.; Arppe, A.; Stacey, A.; and Silfverberg, M. P. 2025. OjibweMorph: An approachable finite-state transducer for Ojibwe. *Language Resources and Evaluation*. Accepted manuscript; preprint.

Hernandez, F.; and Nguyen, V. 2020. The Ubiqus English–Inuktitut System for WMT20. In *Proceedings of the Fifth Conference on Machine Translation*, 213–217. Online: Association for Computational Linguistics.

Joanis, E.; Knowles, R.; Kuhn, R.; Larkin, S.; Littell, P.; Lo, C.-k.; Stewart, D.; and Micher, J. 2020. The Nunavut Hansard Inuktitut–English Parallel Corpus 3.0 with Preliminary Machine Translation Results. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 2562–2572. Marseille, France: European Language Resources Association.

Joshi, P.; Santy, S.; Budhiraja, A.; Bali, K.; and Choudhury, M. 2020. The State and Fate of Linguistic Diversity and Inclusion in the NLP World. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Kirk, D. 2024. Amplifying Inuktitut Voices in the Digital Age with the Power of Technology. https://news.microsoft.com/source/canada/features/uncategorized/amplifying-inuktitut-voices-in-the-digital-age-with-the-power-of-technology/.

Knowles, R.; Stewart, D.; Larkin, S.; and Littell, P. 2020. NRC Systems for the 2020 Inuktitut–English News Translation Task. In *Proceedings of the Fifth Conference on Machine Translation*, 156–170. Online: Association for Computational Linguistics.

Littell, P.; Joanis, E.; Pine, A.; Tessier, M.; Huggins Daines, D.; and Torkornoo, D. 2022. ReadAlong Studio: Practical Zero-Shot Text–Speech Alignment for Indigenous Language Audiobooks. In *Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages*, 23–32. Marseille, France: European Language Resources Association.

Micher, J. 2017. Improving Coverage of an Inuktitut Morphological Analyzer Using a Segmental Recurrent Neural Network. In *Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages (ComputEL-2)*.

Microsoft News Center Canada. 2021. Microsoft Introduces Inuktitut to Microsoft Translator. https://news.microsoft.com/en-ca/2021/01/27/microsoft-introduces-inuktitut-to-microsoft-translator/.

Moshagen, S. N.; Pirinen, F.; Antonsen, L.; Gaup, B.; Mikkelsen, I.; Trosterud, T.; Wiechetek, L.; and Hiovain-Asikainen, K. 2023. The GiellaLT infrastructure: A multilingual infrastructure for rule-based NLP. In *Rule-Based Language Technology*, NEALT Monograph Series.

Nagoudi, E. M. B.; Chen, W.-R.; Abdul-Mageed, M.; and Çavusoglu, H. 2021. IndT5: A Text-to-Text Transformer for 10 Indigenous Languages. *arXiv preprint arXiv:2104.07483*.

NLLB Team; Costa-jussà, M. R.; Cross, J.; Çelebi, O.; Elbayad, M.; Heafield, K.; Ayan, N. F.; Bhosale, S.; Edunov, S.; Fan, A.; Gao, C.; Guzmán, F.; Koehn, P.; Schwenk, H.; et al. 2022. No Language Left Behind: Scaling Human-Centered Machine Translation. *arXiv preprint arXiv:2207.04672*.