



Funded by
the European Union



UNIVERSITY OF
THE FAROE ISLANDS

TrustLLM



UNIVERSITY
OF ICELAND

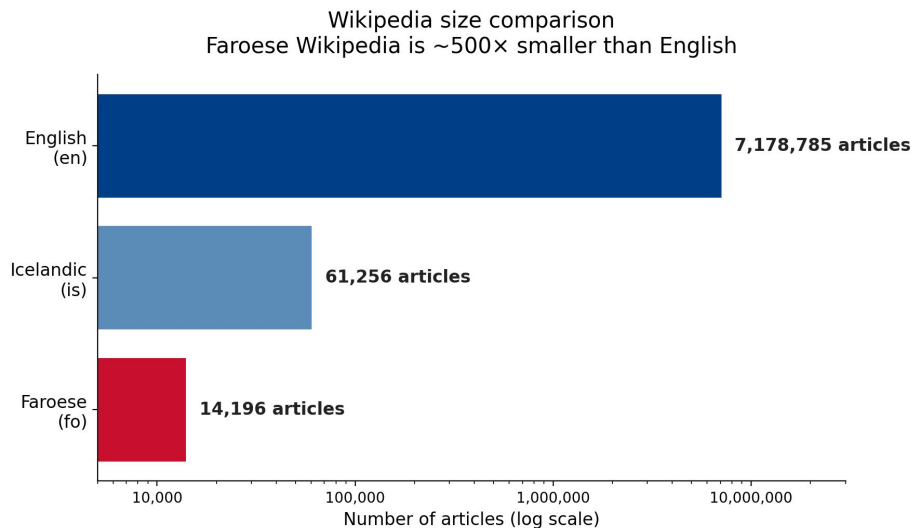
Entity Linking for Faroese Using Large Language Models with Web Search

*Annika Simonsen, Iben Nyholm Debess and Hafsteinn Einarsson, 11th of May 2026
RESOURCEFUL-2026, Palma, Spain*

School of Engineering and Natural Sciences

Introduction

- ~70,000 speakers, primarily in the Faroe Islands
- North Germanic, closely related to Icelandic
- Morphologically rich (4 cases, complex inflection)
- Limited NLP resources: NER model exists, but no entity linker
- Faroese Wikipedia: ~14,196 articles (Feb 2026)



Sources: en.wikipedia.org (May 2026) · is.wikipedia.org (March 2026) · fo.wikipedia.org (Feb 2026)

What's Entity Linking?

- Connect mentions in text → entries in a knowledge base (typically Wikipedia)
- Pipeline traditionally: mention detection → candidate generation → disambiguation
- Each step usually requires language-specific resources and training data
- For low-resource languages: this is a chicken-and-egg problem


Tórshavn er høvuðsstaðurin í Føroyum



 Wikipedia
<https://en.wikipedia.org/wiki/Tórshavn>

Tórshavn

The capital and largest city of the Faroe Islands. It is located in the southern part on the east coast of Streymoy. [Read more](#)




 Wikipedia
https://en.wikipedia.org/wiki/Faroe_Islands

Faroe Islands

The Faroe Islands also known as the Faroes, are an archipelago in the North Atlantic Ocean and a rígsdel ('autonomous territory') of the Kingdom of Denmark. [Read more](#)

Why Faroese is hard

"Herfyri kom flakatrolarin, **Enniberg**, aftur úr **Barentshavinum** við 850 tonsum av flaki." (The factory trawler **Enniberg** came back from the **Barents Sea** with 850 tons of fillet.)

- **Name ambiguity:** "Enniberg" = a fishing vessel here, but also a famous 754m sea cliff



- **Morphological variation:** "Barentshavinum" (dative) → must resolve to "Barentshavið" (nominative)



Research Questions

- Can a frontier LLM with web search perform entity linking for Faroese zero-shot, without task-specific training?
- How does performance vary across entity types?
- Does this approach generalize across models?

Our Approach

One-step entity linking: GPT-5 + integrated web search via OpenRouter

No fine-tuning, no candidate database, no separate disambiguator

Three-tier fallback strategy:

1. Faroese Wikipedia (fo)
2. English Wikipedia (en)
3. Any Wikipedia (any language)

Structured JSON output using PyDantic: "PageTitle >> lang_code" or ""

For example: "Tórshavn >> fo"

Data and Evaluation

- Source: Faroese NER dataset (Snæbjarnarson et al., 2023), Sosialurin news articles
- 3,000 examples → 5,584 entity mentions
- Manual annotation subset: 1,010 examples, 1,647 predictions
- Two annotators, three labels: Correct / Incorrect / Uncertain
- Metrics: Precision, Recall, F1, Cohen's Kappa
- Released as the first gold-standard EL dataset for Faroese

Main Results

Annotator	Precision	Recall	F1
Annotator 1	87.5%	87.3%	0.874
Annotator 2	87.3%	82.8%	0.850

- **67.6% of 5,584 entities received non-empty links**
- **Inter-annotator agreement: 94.5% ($\kappa = 0.706$, excluding "uncertain")**

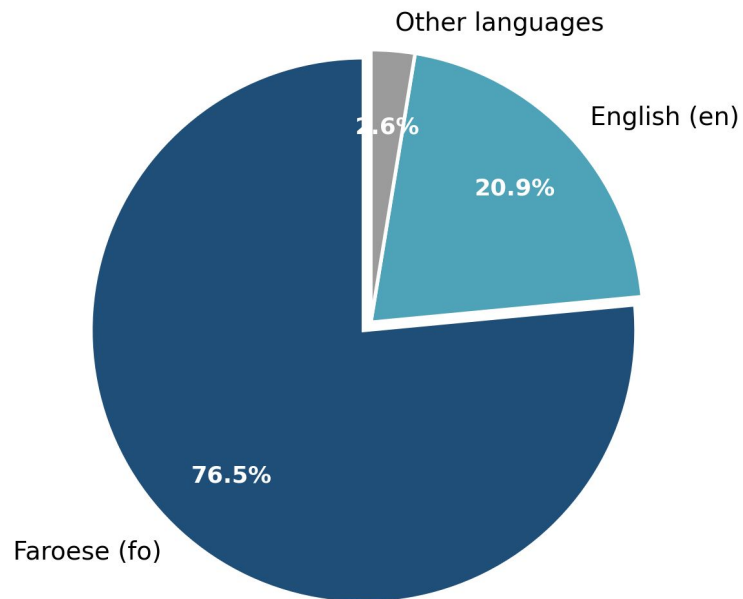
Performance by Entity Type

Annotator	Type	P	R	F1
Annotator 1	PER	85.9%	83.3%	0.846
	LOC	93.5%	94.6%	0.941
	ORG	82.6%	81.6%	0.821
	MISC	61.8%	63.6%	0.627
Annotator 2	PER	87.7%	71.9%	0.790
	LOC	93.3%	92.5%	0.929
	ORG	81.4%	78.0%	0.797
	MISC	61.8%	58.3%	0.600

Table 4: Performance by entity type for each annotator. Type: PER = Person, LOC = Location, ORG = Organization, MISC = Miscellaneous. P = Precision, R = Recall.

Where do the links go?

Wikipedia language distribution of generated links
(N = 3,774 non-empty links)



Other languages: German 1.5% · Danish 0.3% · Norwegian 0.2% · Spanish 0.2% · Icelandic 0.1% · Hungarian 0.1% · French 0.1% · Russian, Swedish <0.1%

Why an LLM?

System	F1 (Ann. 1)	F1 (Ann. 2)
Wikipedia API search	0.573	0.589
GPT-5 + web search	0.874	0.850

- Search baseline finds more candidates (74.7% vs 67.6% coverage)
- But returns plausible-looking wrong pages
- LLM's advantage: **knowing when not to link**, contextual disambiguation
- Biggest gap: Person and Organization entities

Does it generalise?

Model	Link Rate	Precision	Recall	F1
GPT-5	67.6%	87.5%	87.3%	0.874
Gemini 3 Flash	70.5%	78.5%	93.1%	0.852
GPT-5.4 Mini	77.5%	61.2%	93.5%	0.740

- All three models work; $F1 = 0.74 - 0.87$
- Frontier models' edge: higher precision (better at refusing to link)
- Smaller models: high recall but more false links

Qualitative Insights

Successes

✓ Handles orthographic variation: Norra → Noreg, Tobbi → Tórbjörn

✓ Resolves contextual first names: Eivør → Eivør Pálsdóttir

Failures

✗ Sometimes links to inflected forms: Gomlurætt (dative) instead of Gamlarætt

✗ Occasionally links first names to articles about the name itself (e.g., "Lars")

Metonymy: *Landsverkfrøðingurinn* (the role) → *Landsverk* (the org); accepted when context supports it



Takeaways and Future Work

Takeaways

- LLMs + web search make zero-shot EL viable for low-resource languages
- Faroese Wikipedia is small but covers 76.5% of links, which is better than expected
- Main bottleneck: knowledge base coverage, also cost

Future Work

- Distill into a smaller, cheaper model using our annotations as training data
- Nordic-prioritized fallback strategy
- Agentic loop: detect empty pages → retry with nominative form
- Local Faroese knowledge bases (newspapers, registries) to complement Wikipedia

Thank you!

Paper: Look out for the proceedings

Code and Dataset: <https://github.com/haffi112/faroese-entity-linking>

My email: annika@hi.is



Funded by EU Horizon TrustLLM (no. 101135671) and the Strategic Research and Development Programme for Language Technology