Morphosyntactic Structure for Low-Resource Language Translation: Background Research

Alex Kraljic, Christopher Nokes

Written: 11/8/2025, Delivered: 11/13/2025

Research Goals and Questions

- **How does language structure impact machine translation?**
  - Machine translators learn structure (in part) through attention.
  - Low-resource environments rarely have enough data to create efficient machine translators.
    - Not enough data to train the attention mechanism.
  - Morphosyntactic taggers require significantly fewer tokens than machine translators.
    - …but low-resource taggers are less accurate.
  - Training for tagging does not necessarily require tagged data.
  - Structure helps translation in high-resource environments.
- **Can we decrease resources required for translation training by including structural data in word embedding?**

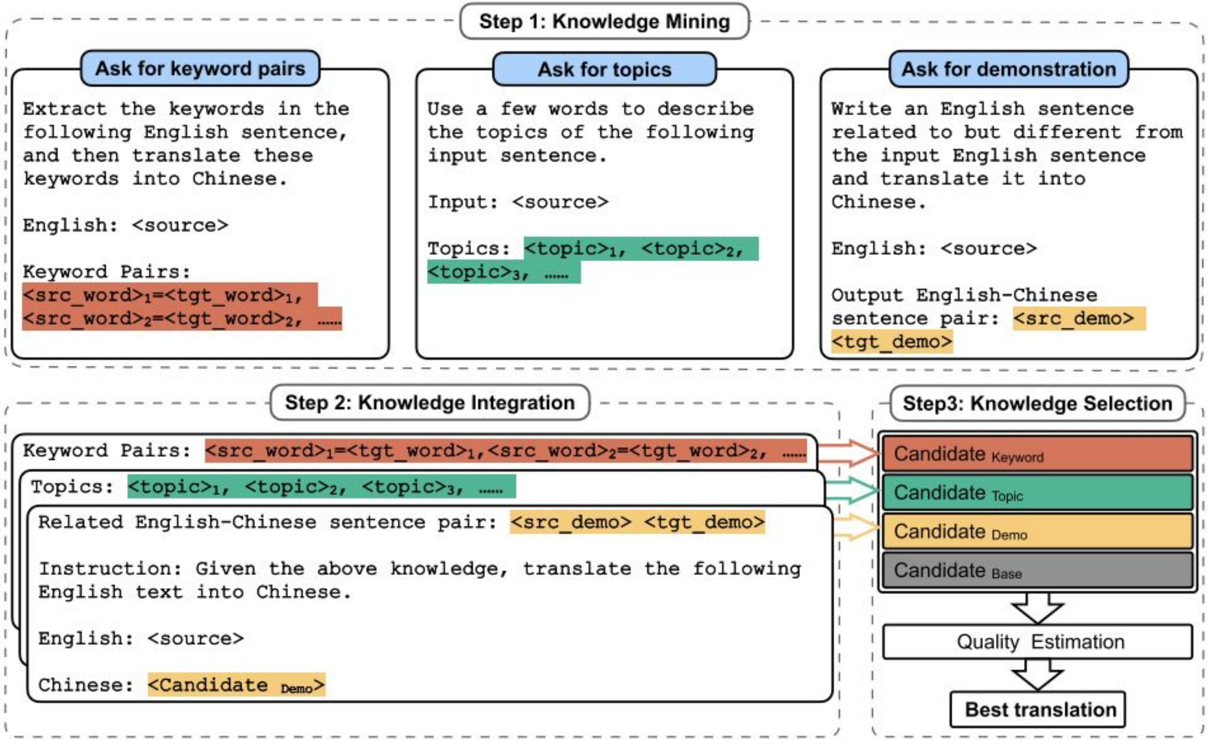
Resource 1: Translation via LLM Reasoning (1/3)

Exploring Human-like Translation Strategy with Large Language Models by Zhiwei He et al.

- **Q: Does an LLM become better at translation when forced to describe structural mechanisms?**
  - A: Yes, significantly!
- **Experiment: break down translation into multiple steps.**
  - Identify source to target pairs for keywords
  - Identify topics in the sentence
  - Perform a similar translation
  - Perform an initial translation
  - Perform a final translation with all of the above knowledge
- **Our takeaway: knowledge of structure helps machine translation**

Resource 1: Translation via LLM Reasoning (2/3)

[*Exploring Human-like Translation Strategy with Large Language Models* by Zhiwei He et al.]



Resource 1: Translation via LLM Reasoning (3/3)

Exploring Human-like Translation Strategy with Large Language Models by Zhiwei He et al.

- **Results: When measured by COMET and BLEURT, successful!**
  - ~30% beneficial - translation with structure better than initial pass.
  - ~50% non-impactful - translation with structure same as initial pass.
  - ~20% detrimental - translation with structure worse than initial pass.
- **How it relates to our work:**
  - This requires an LLM – low-resource languages have nowhere near enough resources to make them.
  - But certain elements here can be mapped to morphosyntactic tags!
    - Keywords are similar to Named Entity Recognition

Resource 2: Approaches for LRLP (1/3)

A Survey on Recent Approaches for Natural Language Processing in Low-Resource Scenarios by Michael Hedderich et al.

- **LRLP: Low-Resource Language Processing**
- **Q: Can LLMs label data quicker than a manual input?**
  - A: Yes, however there are more errors.
- **Experiment: Testing different methods of data labeling, including:**
  - Data augmentation
  - Distant supervision
  - Embeddings and pre-trained LLMs, LLM domain adaptation
  - Multilingual language models and cross-lingual projections
  - Adversarial discriminator and meta-learning
- **Our takeaway: LLMs are capable of self labeling in translation, but it must be used carefully.**

Resource 2: Approaches for LRLP (2/3)

A Survey on Recent Approaches for Natural Language Processing in Low-Resource Scenarios by Michael Hedderich et al.

Group	Task	Yoruba	Hausa	Quechuan	Nahuatl	Estonian
	Num-Speakers	40 mil.	60 mil.	8 mil.	1.7 mil.	1.3 mil.
Text processing	Word segmentation	✓	✓	✓	✓	✓
	Optical character recognition	Hakro et al. (2016)	Hakro et al. (2016)	Hakro et al. (2016)	Hakro et al. (2016)	Hakro et al. (2016)
Morphological analysis	Lemmatization / Stemming	Cotterell et al. (2018)	Cotterell et al. (2018)	Cotterell et al. (2018)	Martínez-Gil et al. (2012)	Cotterell et al. (2018)
	Part-of-Speech tagging	Nivre et al. (2020)	Tukur et al. (2019)	Lozano et al. (2013)	✗	Nivre et al. (2020)
Syntactic analysis	Sentence breaking	✓	✓	✓	✓	✓
	Parsing	Nivre et al. (2020)	✗	Nivre et al. (2020)	✗	Nivre et al. (2020)
Distributional semantics	Word embeddings	FT, BPEmb	FT, BPEmb	FT, BPEmb	FT, BPEmb	FT, BPEmb
	Transformer models	mBERT	XLM-R	✗	✗	mBERT, XLM-R
Lexical semantics	Named entity recognition	Adelani et al. (2020)	Adelani et al. (2020)	Pan et al. (2017)	Pan et al. (2017)	Tkachenko et al. (2013)
	Sentiment analysis	✗	✗	✗	✗	Pajupuu et al. (2016)
Relational semantics	Relationship extraction	✗	✗	✗	✗	✗
	Semantic Role Labelling	Tracey and Strassel (2020)	Tracey and Strassel (2020)	✗	✗	✗
	Semantic Parsing	Nivre et al. (2020)	✗	✗	✗	Nivre et al. (2020)
Discourse	Coreference resolution	✗	✗	✗	✗	Kübler and Zhekova (2016)
	Discourse analysis	✗	✗	✗	✗	
	Textual entailment	Hu et al. (2020)	✗	✗	✗	Hu et al. (2020)
Higher-level NLP	Text summarization	✗	Bashir et al. (2017)	✗	✗	Müürisep and Mutso (2005)
	Dialogue management	✗	✗	✗	✗	✗
	Question answering (QA)	Hu et al. (2020)	✗	✗	✗	Hu et al. (2020)
	SUM	13	10	8	6	15

Table 3: Overview of tasks covered by six different languages. Note that this list is non-exhaustive and due to space reasons we only give one reference per language and task.

Resource 2: Approaches for LRLP (3/3)

A Survey on Recent Approaches for Natural Language Processing in Low-Resource Scenarios by Michael Hedderich et al.

- **Results:**
  - LLMs could most reliable perform two label sets:
    - Word segmentation
    - Sentence break phrasing
- **How it relates to our work:**
  - We can perform some high-level tasks with minimal data
    - Word segmentation
    - Sentence breakdowns
    - Phrasing
  - This data may prove critical for low-resource translation

Resource 3: Performance in Low Resource POS Tagging (1/3)

Weakly Supervised POS Taggers Perform Poorly on Truly Low-Resource Languages by Katharina Kann et al.

- **Q: Can POS tagging be done in low-resource languages?**
  - A: Yes, but it is highly inaccurate and slow, under 50% accuracy.
- **Experiment: Have a variety of LLMs tag different languages.**
  - CHR11
  - GAR13
  - PLA16
  - AMB & AMB+AE
  - FREQ & FREQ+AE
- **Our takeaway: traditional tagging methods may prove challenging in low-resource environments.**

Resource 3: Performance in Low Resource POS Tagging (2/3)

Weakly Supervised POS Taggers Perform Poorly on Truly Low-Resource Languages by Katharina Kann et al.

Language		Treebank Data (test)		UNIMORPH	WIKIDATA+PANLEX	WIKIPEDIA (# tagged)		Embeddings
code	family	sentences	tokens	entries	translations	sentences	tokens	entries
am	AA	1,095	10k	-	2.7k	777	17.9k	10k
be	IE	68	1.3k	-	35.3k	7,385	101.9k	93k
br	IE	888	10.3k	-	12.2k	9,083	112.9k	39k
fo	IE	1,208	10.0k	45.4k	2.9k	9,958	144.6k	12k
hsb	IE	623	10.7k	-	4.6k	1,858	30.2k	10k
hy	IE	514	11.4k	338k	65.1k	3,560	71.4k	47k
kmr	IE	734	10.1k	-	4.6k	3,225	48.3k	24k
lt	IE	55	1.0k	34.1k	38.9k	11,464	117.2k	100k
mr	IE	47	0.4k	-	23.4k	4,886	55.2k	47k
mt	AA	100	2.3k	-	2.1k	2,361	43.9k	16k
bxr	Mo	908	10.0k	-	2.7k	2,308	37.8k	28k
kk	Tu	1,047	10.1k	-	63.5k	12,273	122.4k	100k
ta	Dr	120	2.2k	-	27.1k	5,772	76.2k	100k
te	Dr	146	0.7k	-	28.0k	7,872	90.9k	100k
tl	Au	55	0.2k	-	6.8k	5,871	97.6k	41k
de	IE	1,000	21.3k	179.3k	90.2k	12,162	195.1k	100k
es	IE	1,000	23.3k	382.9k	59.7k	15,209	276.6k	100k
it	IE	1,000	23.7k	509.5k	59.7k	10,254	170.0k	100k
pt	IE	1,000	23.4k	303.9k	47.9k	12,674	195.2k	100k
sv	IE	1,000	19.1k	78.4k	58.8k	10,243	134.5k	100k

Table 1: Resources for our low-resource languages (up) and high-resource languages (down). Language families: Afro-Asiatic (AA), Austronesian (Au), Dravidian (Dr), Indo-European (IE), Mongolic (Mo), and Turkic (Tu).

Resource 3: Performance in Low Resource POS Tagging (3/3)

Weakly Supervised POS Taggers Perform Poorly on Truly Low-Resource Languages by Katharina Kann et al.

- **Results:**
  - GPOS tagging is difficult due to limited resources
  - <50% accuracy in worst cases
- **How it relates to our work:**
  - Traditional tagging methods are ineffective with minimal data
  - Consideration: do we need to take a different tagging approach?
  - Consideration: how accurate does the tagger need to be for morphosyntactic data to aid translation?
  - Tagger accuracy must be measured and tracked

Resource 4: Applied Low-Resource NLP (1/3)

Practical Natural Language Processing for Low-Resource Languages by Benjamin King

- **Q: How can accuracy of LRMs in tagging be increased?**
  - A: Use multiple source and simultaneous target languages
- **Experiment: Testing different ways to raise the accuracy**
  - Increased redundancy
  - Increase the range of syntactic phenomena
- **Our takeaway: By adding fallbacks and running checks on itself, LRMs can reliably be used in language tagging.**

Resource 4: Applied Low-Resource NLP (2/3)

Practical Natural Language Processing for Low-Resource Languages by Benjamin King

Language	Täckström et al.	Single source, single target	Multi source, single target	Single source, multi target	Multi source, Multi target
Danish	77.67	82.55	**85.13***	82.64	83.37*
Dutch	84.28	83.92	**85.25***	84.05	84.35
German	88.16	88.57	**90.45***	88.84	90.02*
Greek	87.57	87.12	**88.82***	86.70	87.01
Italian	86.73*	86.17	**87.75***	85.82	84.54
Portuguese	84.71	88.19	**88.31**	82.19	86.69
Spanish	87.37	87.45	**89.14***	86.93	87.72
Swedish	80.43	80.29	**83.03***	82.43*	82.37*
Average	84.62	85.53	**87.23***	84.95	85.76

Table 7.15: Accuracies of this chapter's methods on each of the target languages. Bolded items represent the highest achieved accuracy for each language. A * indicates that an entry is statistically significantly better than the single-source single-target entry with $p < 0.01$.

Resource 4: Applied Low-Resource NLP (3/3)

Practical Natural Language Processing for Low-Resource Languages by Benjamin King

- **Results:**
  - Improved cross-lingual POS tagging accuracy
  - Statistically significant lower error rate
  - Demonstrated applications in downstream tasks
  - Best case: multiple source languages, one target language
- **How it relates to our work:**
  - Provides methods for improving tagging when lacking resources
  - Gives general targets for accuracy, token count, etc.
  - Solidifies relationship between translation and structure

Models and Datasets

- **We need two, untrained models: a tagger and translator.**
- **Tagger: spaCy!**
  - Has support for a lot of languages.
  - We can use their framework but perform our own training.
- **Translator: Hugging Face!**
  - General transformer structure we can train from scratch.
  - Allows us to build our own input type (word + structure data).
- **Two datasets per language: POS tagged and parallel.**
  - Tagged: Universal Dependencies
  - Parallel: Open Parallel Corpora (OPUS)

Language Selection

- **What languages do we want to use for translation?**
- **We’ll do translations *to* or *from* English.**
- **How low-resource should the language be?**
  - Too many resources? We can artificially make it lower resource by using fewer tokens for training.
  - We need English to language parallel data.
  - We need tagged data for the language.
- **We’re deliberately designing our tool to be language-agnostic.**
  - We only need the parallel and tagged data.
- **So, what languages are we actually using?**
  - English, Croatian, Telugu, and more?

References

- **He, Zhiwei, et al. "Exploring Human-like Translation Strategy with Large Language Models."** Transactions of the Association for Computational Linguistics, vol. 12, 1 Jan. 2024, pp. 229–246, https://doi.org/10.1162/tacl_a_00642. Accessed 30 May 2024.
- **Hedderich, Michael, et al. A Survey on Recent Approaches for Natural Language Processing in Low-Resource Scenarios.** 9 Apr. 2021.
- **Kann, Katharina, et al. "Weakly Supervised POS Taggers Perform Poorly on Truly Low-Resource Languages."** Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, no. 05, 3 Apr. 2020, pp. 8066–8073, ojs.aaai.org/index.php/AAAI/article/view/6317, https://doi.org/10.1609/aaai.v34i05.6317. Accessed 4 Nov. 2025.
- **King, Benjamin. Practical Natural Language Processing for Low-Resource Languages.** 2015.