Cognate and Contact-Induced Transfer Learning for Hamshentsnag: A Low-Resource and Endangered Language

Onur Keleş
Boğaziçi University
Department of Linguistics
onur.keles1@bogazici.edu.tr

Baran Günay
Boğaziçi University
İstanbul Aydın University
baran.gunay@std.bogazici.edu.tr

Berat Doğan
Boğaziçi University
berat.dogan@std.bogazici.edu.tr

Abstract

This study investigates zero-shot and few-shot cross-lingual transfer effects in Part-of-Speech (POS) tagging and Named Entity Recognition (NER) for Hamshentsnag, an endangered Western Armenian dialect. We examine how different source languages, Western Armenian (contact cognate), Eastern Armenian (ancestral cognate), Turkish (substrate or contact-induced), and English (non-cognate), affect the task performance using multilingual BERT and BERTurk. Results show that cognate varieties improved POS tagging by 8% F1, while the substrate source enhanced NER by 15% F1. BERTurk outperformed mBERT on NER but not on POS. We attribute this to task-specific advantages of different source languages. We also used script conversion and phonetic alignment with the target for non-Latin scripts, which alleviated transfer.

Introduction

This study examines cross-lingual transfer from contact and cognate variety languages in Part-of-Speech (POS) and Named Entity Recognition (NER) tagging for a truly low-resource and endangered language Hamshentsnag[1] (hyh). While supervised sequence tagging is a solved problem for high-resource languages (Bohnet et al., 2018), it is indeed difficult for truly low-resource settings with mean accuracies below 50% (Sonkar et al., 2023; Cho et al., 2018; Kann et al., 2020; Malmasi et al., 2022; Choenni et al., 2023), especially in the dearth of available annotated data.

NLP technologies remain limited for underserved communities, and model accuracies in various NLP tasks are significantly lower for languages and cultures that are less represented (Myung et al., 2024). Many available solutions include either continual mixed language pre-training (Liu et al., 2021), using parallel corpora (Ramesh et al., 2022), or employing cross-lingual transfer methods from a higher-resource to a lower-resource language by fine-tuning pre-trained models to increase performance in downstream NLP tasks (Eronen et al., 2023; Cotterell and Duh, 2017). To this end, we have curated a small Hamshentsnag dataset with online resources and working together with the Hemshin community (data elicitation) and employed zero-shot and few-shot cross-lingual transfer by testing two models (i) multilingual BERT (mBERT) (Devlin et al., 2019) and BERT model for Turkish (BERTurk) (Schweter, 2020) for sequence tagging. The source languages were Western Armenian - hyw, Eastern or Standard Armenian - hy; and Standard Modern Turkish - tr), and English (en) that have more resources available (Figure 1). We use the terminology in Table 1 to refer to these languages in the present study. Among the source languages, tr is a substrate to the target; hyw and hy are cognates that share structural similarity with the target, and English (as a reference level for our comparisons) has no contact and little typological similarity.

From a typological background, hy and hyw are distinct dialects of Armenian, but to some degree they are mutually intelligible. hyw has phonological and syntactic differences from hy. hyw retains most of the features of Classical Armenian (Dum-Tragut, 2009), whereas hyw underwent relatively more morpho-phonological and morpho-syntactic simplifications. hyh (the target language) is closest to hyw, while being highly influenced by tr due to prolonged contact. Moreover, the interaction

[1] https://glottolog.org/resource/languoid/id/hams1239

Proceedings of the 1st Workshop on Language Models for Underserved Communities (LM4UC 2025), pages 76–85
May 4, 2025

between the historical hyw and hyh speakers led to potential linguistic exchange or shared features (Khanjian, 2013).

The Hamshentsnag[2] Language

Hamshentsnag (hyh) is considered a dialect of Western Armenian (hyw) (Vaux, 2001), which belongs to the Armenic branch of the Indo-European family. Following the claims (Vaux, 2007) about hyh's typological status, its syntax (Günay et al.) and lexicon, we selected typologically similar languages to transfer knowledge from, which are: hyw, hy, and tr. Our decision behind choosing these three source languages comes from the following features of hyh: the typological landscape of hyh resembles hyw, hy, and tr, in terms of its syntax and morphology. The shared similarities between these three languages are listed below (1), (2). The similarities between the Armenic languages are evident. All of the words in (1-a) to (1-e) in bold are *postpositions*, which are commonly attested in the languages in question (Stevick, 1955). The importance that this carries comes from the ordering in the nominal domain w.r.t. each language, (1) is just one example.

(1)
a. dun-e medan **hedev** — hyh
b. dun ertale **jedk** — hyw
c. tun-e mat'neluc **het'o** — hy
d. ev-e girdikten **sonra** — tr
e. 'after entering the home' — en

Furthermore, the boldfaced morphemes in (2-a) (2-b) (2-c) are the definite (DEF) markers, which are obligatory with proper names. (2-e) is the translation. The boldfaced morpheme in (2-d) is not a definite marker but a genitive (GEN) suffix, it resembles the Hamshentsnag morphology *-i-n* (-GEN-DEF) in terms of its form.

(2)
a. Hasan-i-**n** *u* Ahmed-i-**n**... — hyh
b. Hasmig-**n** *u* Aram-**e**... — hyw
c. Hasmig-**n** *u* Aram-**n**... — hy
d. Hasan-**ın** ve Ahmet-**in**... — tr
e. 'Hasan and Ahmet...' — en

In all three Armenic languages, even the definite marker is subject to the same phonological (3) and morphological (4) constraints (Sigler, 1997):

(3) $/\text{-DEF}/ \rightarrow [e]\ /\ [\text{CONSONANT}]\_\_$

(4) $/\text{-DEF}/ \rightarrow [n]\ /\ \_\_[\text{CLITIC}_{\text{al, u, ...}}]$

Ultimately, the aforementioned observations veered us in selecting these three languages as sources, in addition to English as a reference level.

Related Work

To our knowledge, there is no computational work specifically on hyh. However, there are studies that investigate mBERT's performance on a variety of low-resource languages. Among them, Lauscher et al. (2020) examined languages from 8 different language families on different NLP tasks and found that transfer performance was strongly aligned with the linguistic similarity of the target and source languages. Pires et al. (2019) also showed that mBERT performed surprisingly well in zero-shot transfer for the POS and NER tasks across many languages and even scripts.

Rahimi et al. (2019) proposed two models (one with an unsupervised transfer and another with a supervised transfer setting by using a small set of 100 target sentences) and evaluated them in a NER task. Using only English as a source language in an unsupervised setting often did not transfer well as opposed to the oracle choice of the source language. Furthermore, in their experiments, script mismatch decreased direct transfer.

Similar to the present study, Şaziye Betül Özateş et al. (2025) and Karagöz et al. (2024) evaluated cross-lingual transfer in both mBERT and BERTurk. The authors introduced *OTA-BOUN*, a Universal Dependencies (UD) treebank for historical Turkish, and fine-evaluated mBERT and BERTurk on POS and NER. They reported improvements when combined with Standard Modern Turkish in the training data, alluding to cross-lingual transfer from a higher-source but out-of-domain variety.

However, languages may not be represented equally in multilingual models. Wu and Dredze (2020) tested mBERT on 153 languages in total for POS and NER, and found improvements in the performance when paired with similar languages to the target, although mBERT is claimed to still learn even in the absence of a shared lexicon or domain across languages (Conneau et al., 2020b), with the caveat that models like mBERT should not be employed alone for low-resource languages.

[2] Hamshentsnag has other names as well: *Homshetsi*, *Homshetsma*. We have been advised by the native speakers to use *Hamshentsnag* when referring to it.

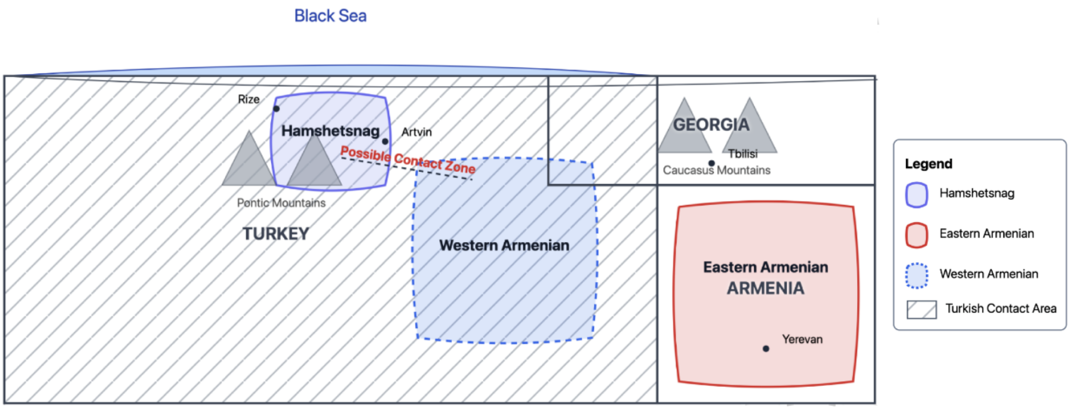

Figure 1: Geographical distribution of Armenian languages in the Caucasus region. The map shows three varieties: Hamshentsnag in northeastern Turkey, Western Armenian's historical speaking area, and Eastern Armenian in modern Armenia. The hatched pattern indicates Turkish linguistic contact areas, while the dotted line between Hamshentsnag and Western Armenian represents a possible historical contact zone.

Term	Definition
Target	The language of interest for which NLP tools are being developed (Hamshentsnag in this study)
Substrate Source (SS)	Language that historically influenced the target language through language contact (e.g., Turkish)
Ancestral Cognate Source (ACS)	Ancestral language that shares a common ancestor with the target language with no contact (e.g., Eastern Armenian)
Contact Cognate Source (CCS)	Language that both influenced the target through contact and shares ancestry with it (e.g., Western Armenian)
Non-Cognate Source (NCS)	Language with no historical contact or a close genetic relationship to the target language (e.g., English)

Table 1: Our Definitions of Language Types

Otherwise, as the authors showed, mBERT performed worse than monolingual models for lower-resource languages. Furthermore, as Artetxe et al. (2020) report, it is not only multilingual models that can learn to generalize to unseen languages, but monolingual models may also transfer at a lexical level and become compatible with mBERT or even perform better. As far as we know, there remains a paucity of research specifically looking at the effects of contact and cognate source languages on the target performance in the context of zero-shot and few-shot cross-lingual transfer from a typological perspective. Also, working together with the community is essential when developing NLP technologies for endangered languages (Liu et al., 2022; Zhang et al., 2022). Therefore, we aim to bridge this gap by investigating how leveraging contact and cognate source languages affects the performance of NLP models specifically for Hamshentsnag and also by collaborating with the Hemshin community and curating relevant linguistic data for few-shot transfer.

Data Resources for Hamshentsnag

Endangered languages come with the cost of the scarcity of data. We alleviated this problem by collecting primary data from four native speakers of the language, who agreed to participate in the data collection process, and written informed consent was obtained from all consultants.[3] Our data collection process was mostly in the form of a Q&A, where the consultants were asked to translate the prepared sentences. Additionally, the consultants were asked to produce sentences about a specific topic. As a second resource, we also utilized a voluntary and nonprofit journal titled *GOR*[4], that aims to preserve the culture, the language, and the history of the Hamshen people. We have benefited from the open-source Hamshen stories that can be found online, which were written in the target language. Lastly, we have benefited from the work of Yenigül (2021), which included in-depth interviews

[3]Data elicitation experiments received ethical approval.

[4]https://gordergi.blogspot.com

with the Hamshen and personal narratives in the target language.

Ultimately, these three approaches increased the number of tokens in our dataset in the following ways: the first approach was tailored towards giving us detailed and crucially more directed naturally produced data, while the second and third approaches were aimed at being efficient with regards to time management in augmenting our dataset, as well as representing the Hamshentsnag language as intended.

Text Normalization and Challenges in Transliteration

The curated data (stored in a text document) were normalized and standardized. The sentences collected for the POS task were further reformatted according to the CoNLL-U format. Because there is no standardized spelling system for the target language, we detected some orthographic inconsistencies (due to speaker variation) and fixed them using Regular Expressions (Regex) together with a researcher who is a native speaker of Hamshentsnag. Since it is spelled using the Latin alphabet, no transliteration was needed.

Additionally, to test multilingual transfer effects, we used Western Armenian (`hyw`) and Eastern Armenian (`hy`) datasets. Using transliteration (i.e., the process of converting text from one writing system to another based on phonetic correspondences) and phonetic transcriptions are known to alleviate cross-lingual transfer (Murikinati et al., 2020; Bharadwaj et al., 2016). For this reason, since these dialects use the Armenian script, we also prepared versions of these datasets which were transliterated into Latin using the `transliterate`[5] package in Python. However, the transliteration outputted by the program significantly differed from the spelling of Hamshentsnag in our corpus given the phonological and orthographic differences between the dialects. Key issues included historical orthographic discrepancies, phonemic variations across dialects, positional allophones, and individual speaker idiosyncrasies. To address these and align the transliteration of `hyw` and `hy` with the target `hyh`, we developed dynamic context-sensitive rules using Regex (see Table 2) by relying on the linguistic judgments and having community validation from our native speaker consultants.

[5]https://pypi.org/project/transliterate

Experimentation

Models The sequence labeling experiments (POS and NER) were implemented using the Flair framework (Akbik et al., 2019). For this task, Google's multilingual BERT (mBERT) (Devlin et al., 2019) and BERTurk (Schweter, 2020) (both of which are cased) were fine-tuned with different training sets, resulting in 9 experiments for POS, and 7 experiments for NER (16 in total) for each model (mmBERT and BERTurk).

mBERT is a multilingual encoder-only model that shares the same architecture with BERT (Devlin et al., 2019) and was trained on 104 different languages. Since `hyh` has close contact with `tr` as illustrated in Figure 1, we also decided to test BERTurk (Schweter, 2020), which is another BERT model trained on a large corpus of Turkish. We also considered XLM-R (Conneau et al., 2020a) but our preliminary experiments showed it underperformed, so we focused on mBERT and BERTurk.

For the fine-tuning, we used the AdamW optimizer with 0.01 weight decay, a learning rate of 5e-5, and a batch size of 32 for a maximum of 15 epochs with early stopping (patience = 5). All experiments were conducted on Google Colab using a Tesla T4 GPU.

POS Data For the POS task, the training datasets include four different languages, as can be seen in Table 3. Our own Hamshentsnag (`hyh`) dataset for POS, described in detail in Section 1, has 373 sentences for the train set, 153 sentences for the development set, and 153 sentences for the test set, all of which were annotated for UPOS by the authors along with native speaker consultants. While the train and development sets come from the same resources (speaker elicitation and open-source Hemshin stories), the test set contains sentences from a different domain (personal experience narratives and dialogues).

The UPOS training data for other higher-resource languages were obtained from Universal Dependency (UD) Treebank datasets. These include Western Armenian (`hyw`), Eastern or Standard Armenian (`hy`), and Turkish (`tr`) to investigate how language contact and cognateness (i.e., typological similarity) contribute to possible multilingual transfer effects. We also trained the model with English (`en`) to test the effect of a high-resource language with no contact and little typological similarity. All testing was conducted only

Armenian Ch.	`transliterate`	Our Transliteration	IPA Transcription
է, ե[†]	ē, e[†]	e, ye[†]	/ɛ/, /je/[†]
ու	ow	u	/u/
ը	ë	ğ	/ə/
շ	š	ş	/ʃ/
չ	č	ç	/tʃ/
ժ	ž	j	/ʒ/
ղ	ġ	ğ	/ʁ/
ռ	ṙ	r	/r/ or /ɾ/[‡]
ջ	ǰ	c/ç	/dʒ/ or /tʃ/[‡]
ո	ò	vo[†] or o	/vo/[†] or /ɔ/
և	ew	yev[†] or ev	/jev/[†] or /ev/
ձ	j	c or ts	/d͡z/ or /t͡s/[‡]
ւ	w	v	/v/

Table 2: Transliteration of the Armenian Script

[†] only word-initially

[‡] in HYW

on the target language (Hamshentsnag or `hyh`).

Dataset	# Sents.	# Tokens
`hyh` (ours)	373	2,394
`hyw` (Yavrumyan et al., 2017)	∼5,000	∼73,000
`hy` (Yavrumyan et al., 2017)	∼2,000	∼34,000
`tr` (Türk et al., 2022)	∼10,000	∼120,000
`en` (Silveira et al., 2014)	∼13,000	∼216,000

Table 3: POS Datasets Used for Training

NER Data We report three languages for the NER task (Table 4). Our own `hyh` developing NER corpus includes a small set of 143 sentences for the training set (all annotated for PERSON (N = 88) and LOCATION (N = 93) entities by the authors under native speaker consultation, consistent with the BIO annotation scheme. Due to the scarcity of open-source data and limitations in linguistic elicitation in Hamshentsnag, other entity types (such as ORGANIZATION) occurred very sparsely and thus were not annotated. Other 46 sentences were curated for the development set, and 115 sentences for the test set (with 109 PER and 58 LOC entities). Like the POS experiment, while the training and development sets in NER came from similar domains and sources (sentences elicited through native speakers and open-source online stories), the test set exclusively included sentences from a different domain and source (personal narrative and dialogues). The NER training set included three higher-resource languages: `hy`, `tr` and `en`, all of which had more than 150K tokens, compared to our own `hyh` corpus with 2K tokens. `hyw` dialect was excluded from NER experiments due to the non-availability of data for this task. Like the previous task, all the testing was done only on the target Hamshentsnag.

Dataset	# Sents.	# Tokens
`hyh` (ours)	143	1785
`hy` (Yavrumyan, 2024)	∼1,000	∼150,000
`tr` (Tür et al., 2003)	∼20,000	∼450,000
`en` (Sang and Meulder, 2003)	∼15,000	∼200,000

Table 4: NER Datasets Used for Training

The descriptions of the model and source combinations for both POS and NER tasks can be found in Table 5.

Experiment Results

POS Each of the 18 models (9 mBERT, 9 BERTurk) were fine-tuned and tested on the target UPOS tags in the test set three times and we report the mean macro-averaged precision, recall, and F1 scores obtained from these experiments. Table 6 illustrates the results for the mBERT models. Zero-shot models (with only `hyw`, `hy`, `tr`, and `en`), we can see that English as a non-contact and non-cognate language performed worse, followed by Turkish (as a substrate or contact-only source). The cognate varieties Eastern and Western Armenian had the best performance. The baseline F1 achieved by the model trained only on our low-resource corpus ($\text{mBERT}_{\texttt{hyh}}$) was 0.63, which could be improved when other contact or cognate languages were added to the training data up to 0.68. The combination of `hyh` and `hyw` resulted in the highest recall (0.70). However, the model trained with both the target and English did not show transfer effects.

The BERTurk models exhibited similar trends in

Model	Description	Train Language
mBERT/BERTURK$_{HYH}$	mBERT/BERTURK fine-tuned only with our own small corpus reported in this study	Target
mBERT/BERTURK$_{HYW}$†	mBERT/BERTURK fine-tuned only with the UD_Western_Armenian-ArmTDP Treebank for POS	CCS
mBERT/BERTURK$_{TR}$	mBERT/BERTURK fine-tuned only with the UD_Turkish-BOUN Treebank for POS and MilliyetNER dataset for NER	SS
mBERT/BERTURK$_{HY}$	mBERT/BERTURK fine-tuned only with the UD_Armenian-ArmTDP Treebank for POS and ArmTDP-NER dataset for NER	ACS
mBERT/BERTURK$_{EN}$	mBERT/BERTURK fine-tuned only with the UD_English-EWT for POS and English CoNLL-2003 dataset for NER	NCS
mBERT/BERTURK$_{hyh+HYW}$†	mBERT/BERTURK fine-tuned with both our hyh corpus and hyw dataset only for POS	Target + CCS
mBERT/BERTURK$_{hyh+TR}$	mBERT/BERTURK fine-tuned with both our hyh corpus and tr datasets for POS and NER	Target + SS
mBERT/BERTURK$_{hyh+HY}$	mBERT/BERTURK fine-tuned with both our hyh corpus and hy datasets for POS and NER	Target + ACS
mBERT/BERTURK$_{hyh+EN}$	mBERT/BERTURK fine-tuned with both our hyh corpus and en datasets for POS and NER	Target + NCS

Table 5: Model Descriptions and Dataset Types Used in the Training Set. CCS: Contact Cognate Source, SS: Substrate Source, ACS: Ancestral Cognate Source, and NCS:Non-Cognate Source.

† These models are only for the POS task since there is no available NER data for hyw.

Model	Precision	Recall	F1
mBERT$_{HYH}$	0.64	0.65	0.63
mBERT$_{HYW}$	0.45	0.37	0.38
mBERT$_{TR}$	0.34	0.27	0.27
mBERT$_{HY}$	0.47	0.35	0.38
mBERT$_{EN}$	0.22	0.22	0.19
mBERT$_{HYH+HYW}$	0.67	**0.70**	0.67
mBERT$_{HYH+TR}$	0.67	0.66	0.66
mBERT$_{HYH+HY}$	**0.69**	0.69	**0.68**
mBERT$_{HYH+EN}$	0.67	0.61	0.63

Table 6: mBERT Results on hyh Test Set for POS

Model	Precision	Recall	F1
BERTURK$_{HYH}$	0.67	0.66	0.64
BERTURK$_{HYW}$	0.43	0.38	0.38
BERTURK$_{TR}$	0.30	0.28	0.26
BERTURK$_{HY}$	0.45	0.34	0.36
BERTURK$_{EN}$	0.21	0.22	0.18
BERTURK$_{hyh+HYW}$	0.67	**0.70**	**0.68**
BERTURK$_{hyh+TR}$	**0.70**	0.69	**0.68**
BERTURK$_{hyh+HY}$	0.67	0.67	0.65
BERTURK$_{hyh+EN}$	0.63	0.59	0.60

Table 7: BERTurk Results on hyh Test Set for POS

POS tagging to the mBERT models, with cognate languages demonstrating superior performance compared to non-cognate languages (Table 7). The baseline model, BERTURK$_{HYH}$, achieved an F1 score of 0.64, which is comparable to the mBERT baseline. When combined with other languages, the BERTurk models showed improvements, with BERTURK$_{hyh+HYW}$ and BERTURK$_{hyh+TR}$ achieving the highest F1 scores of 0.68. Notably, BERTURK$_{hyh+TR}$ also attained the highest precision (0.70), while BERTURK$_{hyh+HYW}$ achieved the highest recall (0.70). However, similar to the mBERT results, the model trained with English (BERTURK$_{hyh+EN}$) showed the least improvement, with an F1 score of 0.60.

NER As in the first experiment, each of the 16 models (8 mBERT, 8 BERTurk) were tested on target NER annotations. The baseline model, mBERT$_{hyh}$, achieved an F1 score of 0.52 (Table 8). Among the zero-shot models, Turkish (mBERT$_{TR}$) achieved the highest precision (0.67) but suffered from low recall (0.31), resulting in a F1 score of 0.35. The model trained on English (mBERT$_{EN}$) performed the worst, with an F1 score of 0.31. When combined with other languages, the mBERT models showed notable improvements: Specifically, mBERT$_{hyh+TR}$ achieved the best performance, with an F1 score of 0.60. In contrast, the model trained with English (mBERT$_{hyh+EN}$) showed limited improvement, achieving an F1 score of 0.47, which is lower than the baseline.

The BERTurk models, on the other hand, demonstrated stronger performance for NER, with the baseline model (BERTURK$_{hyh}$) achieving an F1 score of 0.57, outperforming its mBERT counter-

Model	Precision	Recall	F1
mBERT$_{\text{hyh}}$	0.57	0.48	0.52
mBERT$_{\text{TR}}$	0.67	0.31	0.35
mBERT$_{\text{HY}}$	0.54	0.33	0.41
mBERT$_{\text{EN}}$	0.46	0.25	0.31
mBERT$_{\text{hyh+TR}}$	**0.79**	**0.51**	**0.60**
mBERT$_{\text{hyh+HY}}$	0.65	0.45	0.52
mBERT$_{\text{hyh+EN}}$	0.58	0.41	0.47

Table 8: mBERT Results on hyh Test Set for NER

part (Table 9). Among zero-shot models, Turkish (BERTURK$_{\text{TR}}$) performed better than English, though both fell short of the baseline. Combining the target language with other languages yielded improvements, with BERTURK$_{\text{hyh+TR}}$ achieving the highest F1 score of 0.64. Adding Armenian (BERTURK$_{\text{hyh+HY}}$) also showed competitive results, while English (BERTURK$_{\text{hyh+EN}}$) did not improve the baseline scores.

Taken together, our findings show that leveraging typologically related or contact languages enhanced model performance in sequence tagging for hyh. Cognate varieties (hyw, hy) improved POS tagging by 8% F1, while substrate language (tr) boosted NER by 15% F1. We also observed that BERTurk consistently outperformed mBERT on NER but not in POS. This result perhaps could be attributed to the substrate influence of tr, which shares lexical and cultural overlap with the target. In contrast, POS tagging might depend more on structural cues, where cognate varieties like hy and hyw (more so possibly due to an additional historical contact with the target) perform better due to their syntactic and morphological convergence with the target language. Overall, both experiments highlight the importance of task-specific language selection for cross-lingual transfer in truly low-resource NLP.

Model	Precision	Recall	F1
BERTURK$_{\text{hyh}}$	0.74	0.49	0.57
BERTURK$_{\text{TR}}$	0.49	0.43	0.46
BERTURK$_{\text{HY}}$	0.61	0.38	0.48
BERTURK$_{\text{EN}}$	0.57	0.33	0.40
BERTURK$_{\text{hyh+TR}}$	**0.77**	**0.54**	**0.64**
BERTURK$_{\text{hyh+HY}}$	0.74	0.53	0.62
BERTURK$_{\text{hyh+EN}}$	0.60	0.55	0.57

Table 9: BERTurk Results on hyh Test Set for NER

Effects of Script and Transliteration We also experimented with the impact of script and phonetic transliteration on model performance, focusing specifically on BERTurk. For POS tagging, Eastern (Standard) Armenian using the Armenian script achieved a macro-averaged F1 score of 0.31. When transliterated to Latin using the `transliterate` package in Python, the F1 score improved to 0.33. Further improvement was observed with our custom transliteration alignment method, which achieved an F1 score of 0.36, as reported earlier. Similarly, for NER, the Armenian script yielded an F1 score of 0.41, while Latin transliteration using the `transliterate` package improved the score to 0.46. Our transliteration alignment method achieved the highest F1 score of 0.48. These results demonstrate that script conversion and phonetic alignment enhance model performance, particularly for languages with non-Latin scripts, aligning well with Muller et al. (2021).

Conclusion

This study explored zero-shot and few-shot cross-lingual transfer for part-of-speech (POS) and named entity recognition (NER) tagging in Hamshentsnag, a truly low-resource and endangered language. By leveraging contact and cognate source languages (Western Armenian, Eastern Armenian, and Turkish), we demonstrated that typologically similar languages significantly improve model performance in sequence tagging tasks. Our experiments revealed that cognate languages, particularly Western Armenian, enhanced POS tagging performance, while Turkish, as a substrate language, transferred most in NER. Additionally, BERTurk outperformed mBERT in NER tasks, likely due to the lexical and cultural overlap between Turkish and Hamshentsnag. Overall, these findings underscore the importance of selecting task-specific source languages for cross-lingual transfer, especially in low-resource settings. Furthermore, our work highlights the value of community collaboration and phonetic transliteration in improving model performance for endangered languages, offering a pathway for future research in under-resourced NLP.

Acknowledgments

We thank Metin Bağrıaçık, Ümit Atlamaz, Tunga Güngör, and Şaziye Betül Özateş for their insightful feedback and comments. Furthermore, our heartfelt thanks go to Cengiz Gülcihan, Ayşegül Gülcihan and our other anonymous Hamshentsnag consultants for their support with their language.

Limitations

This study has several limitations that warrant consideration: (i) the dataset for Hamshentsnag remains small due to the lack of open-source online resources and due to working with a relatively small number of language consultants, which inevitably leads to a rather restricted amount of data collection process. This may limit the generalizability of our findings. In addition, (ii) our preliminary hyh dataset at this stage includes sentences from similar domains (mostly stories, personal experiences, and dialogues) and lacks other domains, which might reduce transferability. Furthermore, (iii) the reliance on transliteration for Armenian scripts introduced potential inconsistencies, despite our efforts to align transliterations with native speaker input. Finally, (iv) while BERTurk showed promise, its performance may not extend to other low-resource languages without similar substrate influences since it is a monolingual model.

References

Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. FLAIR: An easy-to-use framework for state-of-the-art NLP. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics (demonstrations)*, pages 54–59.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Akash Bharadwaj, David Mortensen, Chris Dyer, and Jaime Carbonell. 2016. Phonologically aware neural model for named entity recognition in low resource transfer settings. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1462–1472, Austin, Texas. Association for Computational Linguistics.

Bernd Bohnet, Ryan McDonald, Gonçalo Simões, Daniel Andor, Emily Pitler, and Joshua Maynez. 2018. Morphosyntactic tagging with a meta-BiLSTM model over context sensitive token encodings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2642–2652, Melbourne, Australia. Association for Computational Linguistics.

Jaejin Cho, Murali Karthick Baskar, Ruizhi Li, Matthew Wiesner, Sri Harish Mallidi, Nelson Yalta, Martin Karafiát, Shinji Watanabe, and Takaaki Hori. 2018. Multilingual Sequence-to-Sequence Speech Recognition: Architecture, Transfer Learning, and Language Modeling. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 521–527.

Rochelle Choenni, Dan Garrette, and Ekaterina Shutova. 2023. How do languages influence each other? studying cross-lingual data sharing during LM fine-tuning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13244–13257, Singapore. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020a. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Alexis Conneau, Shijie Wu, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. 2020b. Emerging cross-lingual structure in pretrained language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6022–6034, Online. Association for Computational Linguistics.

Ryan Cotterell and Kevin Duh. 2017. Low-resource named entity recognition with cross-lingual, character-level neural conditional random fields. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 91–96, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Şaziye Betül Özateş, Tarık Emre Tıraş, Ece Elif Adak, Berat Doğan, Fatih Burak Karagöz, Efe Eren Genç, and Esma F. Bilgin Taşdemir. 2025. Building foundations for natural language processing of historical Turkish: Resources and models. *Preprint*, arXiv:2501.04828.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Jasmine Dum-Tragut. 2009. *Armenian: Modern Eastern Armenian*, volume 14 of *London Oriental and African Language Library*. John Benjamins Publishing Company, Amsterdam/Philadelphia.

Juuso Eronen, Michal Ptaszynski, and Fumito Masui. 2023. Zero-shot cross-lingual transfer language selection using linguistic similarity. *Information Processing Management*, 60(3):103250.

Baran Günay, Ümit Atlamaz, and Ömer Demirok. Single conjunct agreement in homshetsma. In *Proceedings of the 55th Annual Meeting of the North East Linguistic Society (NELS 55)*. To appear.

Katharina Kann, Ophélie Lacroix, and Anders Søgaard. 2020. Weakly supervised pos taggers perform poorly on truly low-resource languages. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8066–8073. Issue: 05.

Fatih Karagöz, Berat Doğan, and Şaziye Betül Özateş. 2024. Towards a clean text corpus for Ottoman Turkish. In *Proceedings of the First Workshop on Natural Language Processing for Turkic Languages (SIGTURK 2024)*, pages 62–70, Bangkok, Thailand and Online. Association for Computational Linguistics.

Hrayr Khanjian. 2013. *(Negative) concord and head directionality in Western Armenian*. Ph.D. thesis, Massachusetts Institute of Technology.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From zero to hero: On the limitations of zero-shot language transfer with multilingual Transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4483–4499, Online. Association for Computational Linguistics.

Zihan Liu, Genta Indra Winata, and Pascale Fung. 2021. Continual mixed-language pre-training for extremely low-resource neural machine translation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2706–2718, Online. Association for Computational Linguistics.

Zoey Liu, Crystal Richardson, Richard Hatcher, and Emily Prud'hommeaux. 2022. Not always about you: Prioritizing community needs when developing endangered language technology. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3933–3944, Dublin, Ireland. Association for Computational Linguistics.

Shervin Malmasi, Anjie Fang, Besnik Fetahu, Sudipta Kar, and Oleg Rokhlenko. 2022. MultiCoNER: A Large-scale Multilingual Dataset for Complex Named Entity Recognition. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3798–3809, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Benjamin Muller, Antonios Anastasopoulos, Benoît Sagot, and Djamé Seddah. 2021. When being unseen from mBERT is just the beginning: Handling new languages with multilingual language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 448–462, Online. Association for Computational Linguistics.

Nikitha Murikinati, Antonios Anastasopoulos, and Graham Neubig. 2020. Transliteration for cross-lingual morphological inflection. In *Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 189–197, Online. Association for Computational Linguistics.

Junho Myung, Nayeon Lee, Yi Zhou, Jiho Jin, Rifki Afina Putri, Dimosthenis Antypas, Hsuvas Borkakoty, Eunsu Kim, Carla Perez-Almendros, Abinew Ali Ayele, Víctor Gutiérrez-Basulto, Yazmín Ibáñez García, Hwaran Lee, Shamsuddeen Hassan Muhammad, Kiwoong Park, Anar Sabuhi Rzayev, Nina White, Seid Muhie Yimam, Mohammad Taher Pilehvar, Nedjma Ousidhoum, Jose Camacho-Collados, and Alice Oh. 2024. Blend: A benchmark for llms on everyday knowledge in diverse cultures and languages. In *Advances in Neural Information Processing Systems*, volume 37, pages 78104–78146. Curran Associates, Inc.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Afshin Rahimi, Yuan Li, and Trevor Cohn. 2019. Massively multilingual transfer for NER. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 151–164, Florence, Italy. Association for Computational Linguistics.

Gowtham Ramesh, Sumanth Doddapaneni, Aravinth Bheemaraj, Mayank Jobanputra, Raghavan AK, Ajitesh Sharma, Sujit Sahoo, Harshita Diddee, Mahalakshmi J, Divyanshu Kakwani, Navneet Kumar, Aswin Pradeep, Srihari Nagaraj, Kumar Deepak, Vivek Raghavan, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh Shantadevi Khapra. 2022. Samanantar: The largest publicly available parallel corpora collection for 11 Indic languages. *Transactions of the Association for Computational Linguistics*, 10:145–162.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Stefan Schweter. 2020. BERTurk-BERT models for Turkish. *Zenodo*, 2020:3770924.

Michele Sigler. 1997. *Specificity and agreement in standard Western Armenian*. Ph.D. thesis, Massachusetts Institute of Technology.

Natalia Silveira, Timothy Dozat, Marie-Catherine de Marneffe, Samuel Bowman, Miriam Connor, John Bauer, and Chris Manning. 2014. A Gold Standard Dependency Corpus for English. In *Proceedings* *of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2897–2904, Reykjavik, Iceland. European Language Resources Association (ELRA).

Shashank Sonkar, Zichao Wang, and Richard Baraniuk. 2023. MANER: Mask augmented named entity recognition for extreme low-resource languages. In *Proceedings of The Fourth Workshop on Simple and Efficient Natural Language Processing (SustaiNLP)*, pages 219–226, Toronto, Canada (Hybrid). Association for Computational Linguistics.

Earl Wilson Stevick. 1955. *Syntax of Colloquial East Armenian*. Cornell University.

Gökhan Tür, Dilek Hakkani-Tür, and Kemal Oflazer. 2003. A statistical information extraction system for turkish. *Natural Language Engineering*, 9(2):181–210.

Utku Türk, Furkan Atmaca, Şaziye Betül Özateş, Gözde Berk, Seyyit Talha Bedir, Abdullatif Köksal, Balkız Öztürk Başaran, Tunga Güngör, and Arzucan Özgür. 2022. Resources for Turkish dependency parsing: Introducing the BOUN treebank and the BoAT annotation tool. *Language Resources and Evaluation*, pages 1–49. Publisher: Springer.

Bert Vaux. 2001. Hemshinli: The forgotten black sea Armenians. *Journal of Armenian studies*, 6(2):47–71.

Bert Vaux. 2007. Homshetsma: The language of the armenians of Hamshen. In *The Hemshin*, pages 257–278. Routledge.

Shijie Wu and Mark Dredze. 2020. Are all languages created equal in multilingual BERT? In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 120–130, Online. Association for Computational Linguistics.

Marat M. Yavrumyan. 2024. ArmTDP-NER: Named Entity Corpus of Modern Eastern Armenian.

Marat M. Yavrumyan, Hrant Khachatrian, Anna Danielyan, and Gor Arakelyan. 2017. ArmTDP: Eastern Armenian treebank and dependency parser. In *XI International Conference on Armenian Linguistics, Abstracts. Yerevan*.

Anıl Yenigül. 2021. *The role of translation in the efforts for the survival of disappearing languages in a globalized world: The case of hemshin*. Ph.D. thesis, Yıldız Technical University.

Shiyue Zhang, Ben Frey, and Mohit Bansal. 2022. How can NLP help revitalize endangered languages? a case study and roadmap for the Cherokee language. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1529–1541, Dublin, Ireland. Association for Computational Linguistics.