



THE UNIVERSITY OF
CHICAGO



From Middle Korean to Altaic-type languages

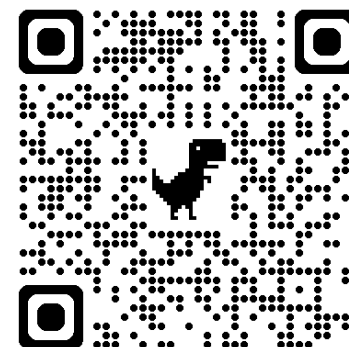
A BiLSTM-based Toolkit for Low-Resource Morphological Analysis

Jaehong Shim

(The University of Chicago, <https://jaehongshim.github.io>)

Sangah Lee

(Seoul National University, <https://sanajlee.github.io>)



Introduction

Processing morphologically complex low-resource languages

- The absence of large-scale training corpora
- Difficulty of leveraging contextual information to resolve morphological syncretism
- Flexible word order

The absence of large-scale training corpora

- The processing capabilities of LLMs rely on very large amounts of training data (Kaplan et al. 2020; Hoffmann et al. 2022; Zhang et al. 2024)

Difficulty of leveraging contextual information to resolve morphological syncretism

- Causative/passive syncretism in Manchu (Gorelova 2002: 248, gloss added by the author)

(1) *wa-bu-*

die-CAUS/PASS

‘to order someone to kill’ or ‘to be killed’

- Dative/genitive syncretism in Middle Korean

(2) *cip-uy*

house-DAT/GEN

‘to a house’ or ‘of the house’

Flexible word order and syncretism

- Flexible word order makes it difficult to identify crucial information to disambiguate syncretism

(3) Causative/passive syncretism in Korean

a. *Cip-eyse* *Chelswu-ka* *po-i-ess-ta.*

house-ABL C.-NOM see-PASS-PST-DECL

‘Cheolsu was visible from the house.’

b. *Chelswu-ka* *Uysa-eykey* *thinwun-ul* *po-i-ess-ta.*

C.-NOM Doctor-DAT corn-ACC see-CAUS-PST-DECL

‘Cheolsu showed his corn to the doctor.’

Flexible word order and syncretism

- Flexible word order makes it difficult to identify crucial information to disambiguate syncretism

(4) Word order change

a.	<i>Chelswu-ka</i>	<i><u>thinwun-ul</u></i>	<i>Uysa-eykey</i>	<i>po-i-ess-ta.</i>
	C.-NOM	corn-ACC	Doctor-DAT	see-CAUS-PST-DECL
	‘Cheolsu showed his corn to the doctor.’			
b.	<i><u>thinwun-ul</u></i>	<i>Chelswu-ka</i>	<i>Uysa-eykey</i>	<i>po-i-ess-ta.</i>
	corn-ACC	C.-NOM	Doctor-DAT	see-CAUS-PST-DECL
c.	<i>Chelswu-ka</i>	<i>uysa-eykey</i>	<i>po-i-ess-ta</i>	<i><u>thinwun-ul.</u></i>
	C.-NOM	doctor-DAT	see-CAUS-PST-DECL	corn-ACC
	...			

- Contextual information should be considered at least at the clausal level.

Successful morphological analysis

- Parsing in the right way
- Disambiguation of syncretism
- Comprehensible integration of contextual information



If only we had
enough time...



One last obstacle: Too many raw texts without appropriate annotation

- Annotation bottleneck (cf. Transcription bottleneck, Chelliah 2001:76)

```
718 <sent type="main" lang="kor" page="25a-25b" n="6">須達이 깃거 象에 金을 시러 여든 塊 싸해 죽자히 다 싣오</sent>(!--頃 ^ 峙-->
719
720 <sent type="anno" lang="kor" page="25b" n="1">頃은 온 畝 |니 혼 畝 |二百 마흔 步 |라</sent>
721 <sent type="main" lang="chi" page="25b" n="2">殘有少地,須達思惟何藏金足,不多不少當足滿之.</sent>
722 <sent type="main" lang="kor" page="25b" n="2">아니한 디 온 다 싣갓거늘 須達이 잔죽고 수랑후더니</sent>
723 <sent type="main" lang="chi" page="25b" n="3">祇陀問言嫌貴置之.</sent>
724 <sent type="main" lang="kor" page="25b" n="3">太子 | 무로디 앓가본 프디 잇느니여</sent>
725 <sent type="main" lang="chi" page="25b" n="4">答言不也.</sent>
726 <sent type="main" lang="kor" page="25b" n="4">對答호디 그리 아니라</sent>
727 <sent type="main" lang="chi" page="25b" n="5">自念金藏何者,可足當補滿之.</sent>
728 <sent type="main" lang="kor" page="25b" n="5">내 수랑호디 어누 藏스 金이와 마치 싣이려뇨 후노이다</sent>
729 <sent type="main" lang="chi" page="25b" n="6">祇陀念言.</sent>
730 <sent type="main" lang="kor" page="25b" n="6">太子 | 녀코디</sent>
731 <sent type="main" lang="chi" page="25b" n="7">佛必大德.</sent>
732 <sent type="main" lang="kor" page="25b" n="7">부텃 德이 至極하사샤</sent>
733 <sent type="main" lang="chi" page="25b-26a" n="8">乃使斯人轉賣乃爾.</sent>
734 <sent type="main" lang="kor" page="25b-26a" n="8">이 사루미 보비를 더라도록 아니 잇기돏다 후야</sent>(!--비^를-->
735
736 <sent type="main" lang="chi" page="26a" n="1">教齊是止</sent>
737 <sent type="main" lang="kor" page="26a" n="1">須達이드려 닐오디</sent>
738 <sent type="main" lang="chi" page="26a" n="2">勿更出金.</sent>
739 <sent type="main" lang="kor" page="26a" n="2">金을 더 내디 말라</sent>
740 <sent type="main" lang="chi" page="26a" n="3">園地屬卿樹木屬我.</sent>
741 <sent type="main" lang="kor" page="26a" n="3">싸온 그딴 모기 두고 남그람 내 모기 두어</sent>
742 <sent type="main" lang="chi" page="26a" n="4">自起門屋上佛,共立精舍.</sent>
743 <sent type="main" lang="kor" page="26a" n="4">둘히 어우러 精舍 興그라 부텃고 받주보리라</sent>
744 <sent type="main" lang="chi" page="26a" n="5">須達歡喜即然可之.</sent>
745 <sent type="main" lang="kor" page="26a" n="5">須達이 깃거 지의 도라가</sent>
746 <sent type="main" lang="chi" page="26a" n="6">便即歸家當施功作.</sent>
747 <sent type="main" lang="kor" page="26a" n="6">精舍 지을 이즈 磨鍊후더니</sent>
748 <sent type="main" lang="chi" page="26a" n="7">六師聞之往白國王.</sent>
749 <sent type="main" lang="kor" page="26a" n="7">그 나랏 六師 | 듣고 王의 勅호디</sent>
750 <sent type="anno" lang="kor" page="26a" n="8">六師는 外道이 스승 여스시라</sent>
751 <sent type="main" lang="chi" page="26a-26b" n="9">長者須達買祇陀園.</sent>
752 <sent type="main" lang="kor" page="26a-26b" n="9">長者 須達이 祇陀太子스 東山을 사야</sent>(!--長^者-->
```

× n ($n \geq$ a considerably big number
you might not want to imagine)

Our proposal

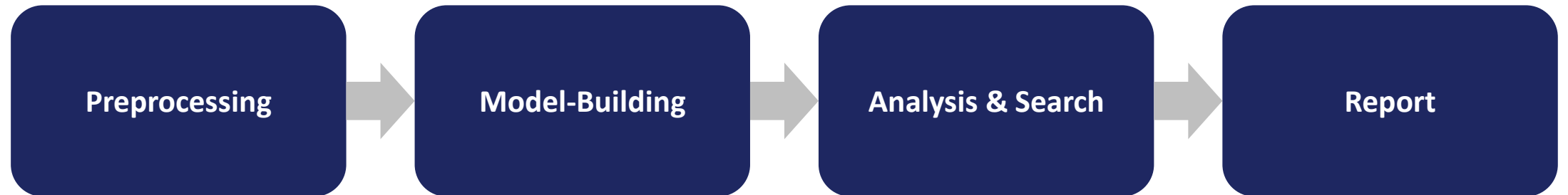
- A Bidirectional Long Short-Term Memory (BiLSTM)-based tool for researchers to build their own morphological-analysis model, from minimal data
 - Researcher's own annotation + BiLSTM model training

From a researcher's own annotation to a fully tagged corpus



The tool in a nutshell

This tool aims to help researchers build their own morpheme-tagging model



We'll walk through each of these stages in detail.

Example implementation

15th c. Middle Korean

Middle Korean

- An ancestral variety of Modern Korean (MoK) spoken from the 10th century to the 15th century.
- Reliable materials are attested from the mid-15th century.



Middle Korean material (1447 *Sŏkpo sangjŏl* 釋譜祥節 11:16)

NIKL Historical Korean Corpora 2023

- Historical Korean corpora from the 15th century to the early 20th century compiled by the National Institute of Korean Language (NIKL).
- XML (UTF-8 encoding)
- Consisting of 1,153 files

NIKL Historical Korean Corpora 2023

Original sentence lines paired with
corresponding metadata

```
<?xml version="1.0" encoding="UTF-8"?>

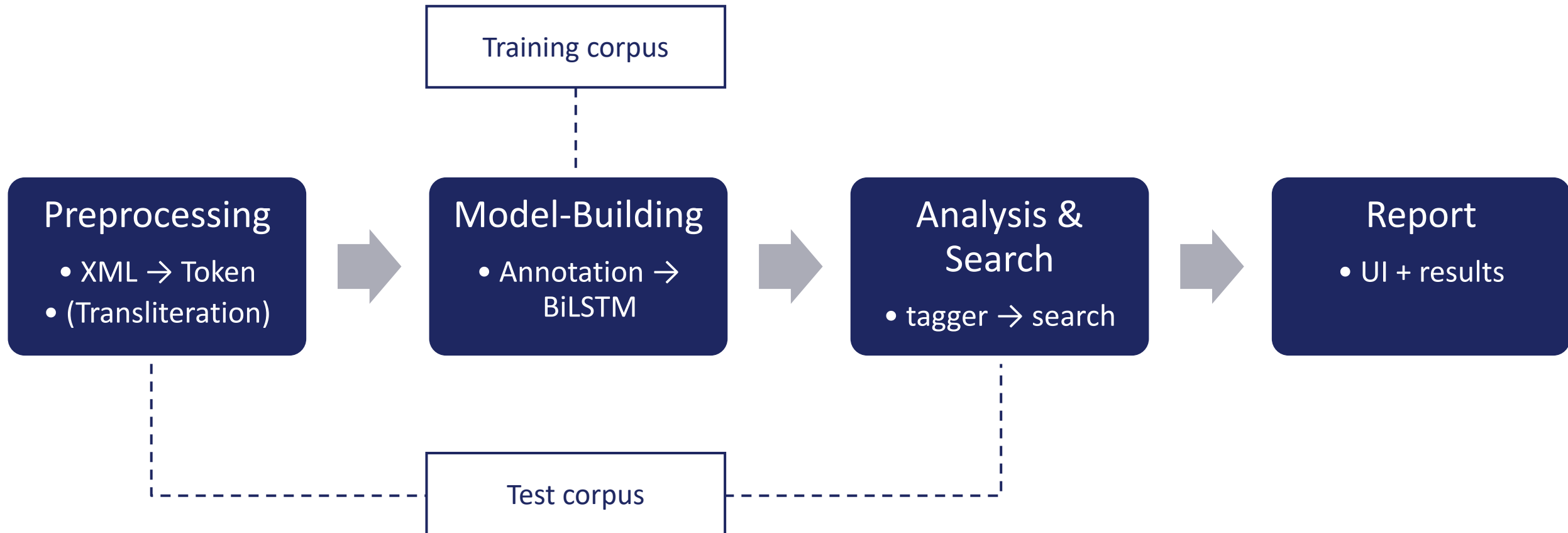
<doc>
<teiHeader>
<titleStmt>
<title lang="kor">석보상절</title>
<title lang="chi">釋譜詳節</title>
<volume n="6"/>
<author></author>
<date>1447</date>
</titleStmt>

...

<sent type="main" lang="kor" page="47a" n="1">여툼 히 니르시고</sent>
<sent type="main" lang="kor" page="47a" n="2">버거 스물흔 히 수시에 般若를 니르시니라</sent>
<sent type="anno" lang="kor" page="47a" n="3">般若 처섬 니루샤미</sent>
<sent type="anno" lang="kor" page="47a" n="4">부텃 나히 쉬너러시니</sent>
<sent type="anno" lang="kor" page="47a" n="5">穆王 7 스물네찰 히 쏫卯 1라</sent>

...
```

The Pipeline



Preprocessing: From XML to Token

Corresponding fields are highlighted in the same color across both panels

Corpus

```
1264 <sent type="main" lang="kor" page="46b" n="4">오직 魔王이 世尊의 술보디</sent>
1265 <sent type="main" lang="kor" page="46b" n="5">瞿曇아 나눈 一切 衆生이 다
부테 두외야</sent>
1266 <sent type="main" lang="kor" page="46b" n="6">衆生이 업거샤 菩提心을 發호리라
후더라</sent>
1267 <sent type="anno" lang="kor" page="46b" n="7">大集經 니루샤미 부텃 나히
마순다수시러시니</sent>
1268 <sent type="anno" lang="kor" page="46b-47a" n="8">穆王 열아흔샷 히 戊戌
이라</sent><!--王 ^ 열-->
1269 <mark type="section_begin">○</mark>
1270
1271 <sent type="main" lang="kor" page="47a" n="1">여둘 히 니르시고</sent>
1272 <sent type="main" lang="kor" page="47a" n="2">버거 스를흔 히 수시에 般若를
니르시니라</sent>
1273 <sent type="anno" lang="kor" page="47a" n="3">般若 처엄 니루샤미</sent>
1274 <sent type="anno" lang="kor" page="47a" n="4">부텃 나히 쉬니러시니</sent>
1275 <sent type="anno" lang="kor" page="47a" n="5">穆王 그 스를네샷 히 쑤卯 이라</
sent>
1276
1277 <sent type="title" lang="chi" page="47a" n="6">釋譜詳節 第六</sent>
```

Token

Token (

...

source_id = 1447_석보상절6:47a:4:kor,

token_index = 3,

lang = kor,

unicode_form = 쉬니러시니,

yale = swinilesini,

is_note = anno,

sent_type = anno,

context = 부텃 나히 <<쉬니러시니>> ,

...

)

Annotation

The annotation mode enables researchers to manually provide a correct morphological analysis for subsequent model training.

- Terminal-based annotation
- Gold JSONL

```
[PROGRESS] 2463/3575 tokens annotated (68.9%) | 1112 remaining
[annotation] 1447_석보상절6:27a:7:kor [D:\Corpus\NIKL\NIKL_Historical Korean Corpus 2023 v2.0\HXRW2320000612.xml]
[Token]      호논고
[LANGUAGE]   kor
[CONTEXT]    므숯 이룰 겻고오려 <<호논고>>
[annotation] No candidates for 호논고 (hononkwo). (m=manual / s=skip / q=quit / press enter=step-by-step annotation) >
Please parse the token for further morpheme tagging: ho no nkwo

Please label ho with an appropriate label: ho/V/LEM
Please label no with an appropriate label: no/PRS
Please label nkwo with an appropriate label: nkwo/Q
```

```
MidKriRegexTool > data > annotation > {} annotation_15c.jsonl
2462 {"period": "15c", "source_id":
      "1447_석보상절6:27a:7:kor",
      "token_index": 4, "token": "호논고",
      "gold_morph": "ho/V/LEM no/PRS nkwo/Q",
      "sent_type": "main", "context": "므숯
      이룰 겻고오려 호논고", "prev_context":
      "분별 말라 六師의 무리 閻浮提에
      구독호야도 내 바렛 혼 터리를 묻
      무으리니", "next_context": "제 홀
      양으로 후게 후라"}
```

Candidate generator: facilitating annotation

The candidate generator automatically extracts possible suffixes from the gold forms in the Gold JSONL file and suggests them as candidates during annotation.

```
[PROGRESS] 2510/3575 tokens annotated (70.2%) | 1065 remaining
[annotation] 1447_석보상절 6:27b:7:kor [D:\Corpus\NIKL\NIKL_Hist
orical Korean Corpus 2023 v2.0\HXRW2320000612.xml]
[Token]      티니
[LANGUAGE]   kor
[CONTEXT]    그 날 다두라 숲 부플 <<티니>>
[CANDIDATES] 1. thi/V/LEM-ni/CONN
              2. thi/V/LEM-n/ADN.NFUT-i/N.DEP
[annotation] What is the optimal candidate for 티니?
(1-2 to select / s=skip / m=manual input / q=quit / press enter
=step-by-step annotation) >1
```

- Input: 티니 **thini**
- Gold *i* → Lemma **thi/V/LEM-ke/KE-nuᄡ/CONN**
- Gold *j* → Suffix 1 **na/V/LEM-ni/CONN**
- Gold *k* → Suffix 2 **sa/V/LEM-n/ADN.NFUT-i/N.DEP**
- Candidate 1: **thi/V/LEM-ni/CONN**
- Candidate 1: **thi/V/LEM-n/ADN.NFUT-i/N.DEP**

Training the BiLSTM model

- The BiLSTM module trains a morphological analysis model on annotated data.
- Forward and backward passes integrate left/right context around each Token.
- Output: a trained model consumed by the tagger in the analysis pipeline.



BiLSTM reranking in action

(5) Previous context: *SYWUTTALQ-i nilwotoy: “nilosyan*
Sudatta-NOM said: what.you.said

Target: *yang-olwo*
DEP.N-INS

Subsequent context: *hwoli.ngita.” THAYCO-i nilwotoy*
will.do prince-NOM said

‘Sudatta said: “I will do as you said.
Then, the prince said: ...”

(1447 *Sökpo sangjöl* 釋譜祥節 6:24b)

BiLSTM Reranking in Action

Illustrative example — how ± 3 -token context resolves a case-marking ambiguity in "양^o로" (yangolwo)



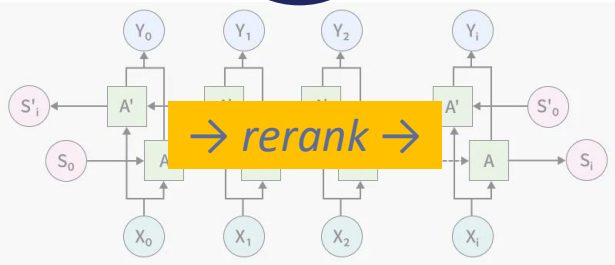
Before BiLSTM

rule-based candidate order (no context)

1. yang/N.DEP/LEM-olwo/ALL
★ selected
2. yang/N/LEM-olwo/ALL
3. yang/N.DEP/LEM-olwo/INS
★ correct analysis
4. yang/N/LEM-olwo/INS

September 23, 2026

BiLSTM



After BiLSTM

reranked using ± 3 -token context

1. yang/N.DEP/LEM-olwo/INS
★ selected
2. yang/N.DEP/LEM-olwo/ALL
3. yang/N/LEM-olwo/ALL
4. yang/N/LEM-olwo/INS

Search

Once the model returns a corpus tagged with a morphological analysis, researchers can use it for either a more precise regular-expression-based search or a tag-based search.

Searching for the Middle Korean object honorific suffix -습 -*sop* with the pattern `[czs]o[pw]`

```
1447_석보상절 6:10b:4:kor 4 main [D:\Corpus\NIKL\NIKL_Historical Korean Corpus 2023_v2.0\HXRW2320000612.xml]
      [TOKEN]          從 후 습 디
      [TAGGED-FORM]     從 ho/V.CH/LEM-zop/HON.OBJ-ti/CONN
      [MATCHED-PART]    zop
      [CONTEXT]         부 테 조 로 니 루 샤 도 <<從 후 습 디>> 아 니 후 더 니
1447_석보상절 6:15b:4:kor 2 main [D:\Corpus\NIKL\NIKL_Historical Korean Corpus 2023_v2.0\HXRW2320000612.xml]
      [TOKEN]          연 조 고
      [TAGGED-FORM]     yetcop/V.HON/LEM-kwo/CONN
      [MATCHED-PART]    cop
      [CONTEXT]         말 미 <<연 조 고>> 천 량 만 히 시 러 王 舍 城 으 로 가 며
1447_석보상절 6:20b:2:kor 2 main [D:\Corpus\NIKL\NIKL_Historical Korean Corpus 2023_v2.0\HXRW2320000612.xml]
      [TOKEN]          부 라 습 고
      [TAGGED-FORM]     pola/V/LEM-zop/HON.OBJ-kwo/CONN
      [MATCHED-PART]    zop
      [CONTEXT]         須 達 이 <<부 라 습 고>> 몬 내 과 후 수 바 호 더
```

Current Progress

- Training data selection and collection complete
- Annotation scheme finalized after outlining the overall model structure
- Data annotation currently underway per the finalized scheme

Current Challenges

- Data scarcity — even within the same period, larger and more varied data is preferable, since it's unclear what test data may come up later.
- Annotation is difficult since we do not have native-speaker intuition — some degree of compromise/acceptance is needed.
- False negative → A fallback stage that shows all the unanalyzable tokens

Toward cross-linguistic application

Toward cross-linguistic application

- Expected to extend well to languages typologically similar to Korean
- Candidate targets: Tungusic, Turkic, Mongolic, and Japonic — languages previously described as Altaic-type, head-final
- Universal Dependencies (UD) format export planned, to ensure interoperability with existing language resources
- Long-term goal: extend beyond typologically similar languages to low-resource languages more broadly

Thank you! Any questions?

Selected references

- Chelliah, Shobhana L. 2021. *Why Language Documentation Matters*. Cham: Springer.
- Gao, Yuan, Ruili Wang, and Feng Hou. 2024. "How to Design Translation Prompts for ChatGPT: An Empirical Study." *Proceedings of the 6th ACM International Conference on Multimedia in Asia Workshops*, December 3, 2024, 1–7.
- Gorelova, Liliya M. 2002. *Manchu Grammar*. Chams: Brill.
- Hoffmann, Jordan, Sebastian Borgeaud, Arthur Mensch, et al. 2022. "Training Compute-Optimal Large Language Models." arXiv:2203.15556. Preprint, arXiv, March 29.
- Kaplan, Jared, Sam McCandlish, Tom Henighan, et al. 2020. "Scaling Laws for Neural Language Models." arXiv:2001.08361. Preprint, arXiv, January 23.
- Ko, Yong-Kun. *Standard Middle Korean Grammar*. 4th edition. Jipmoondang, 2020.
- National Institute of Korean Language. 2023. *NIKL Historical Korean Corpus 2023 (v2.0)*.
<https://kli.Korean.go.kr/corpus>
- Zhang, Kexun, Yee Man Choi, Zhenqiao Song, Taiqi He, William Yang Wang, and Lei Li. 2024. "Hire a Linguist!: Learning Endangered Languages in LLMs with in-Context Linguistic Descriptions." *Findings of the Association for Computational Linguistics: ACL 2024*, 15654–69.
<https://aclanthology.org/2024.findings-acl.925/>.