NLP Models for Field Linguistic Annotations in Computational Language Documentation

ILFC Seminar

Shu OKABE

16 April 2026

About me

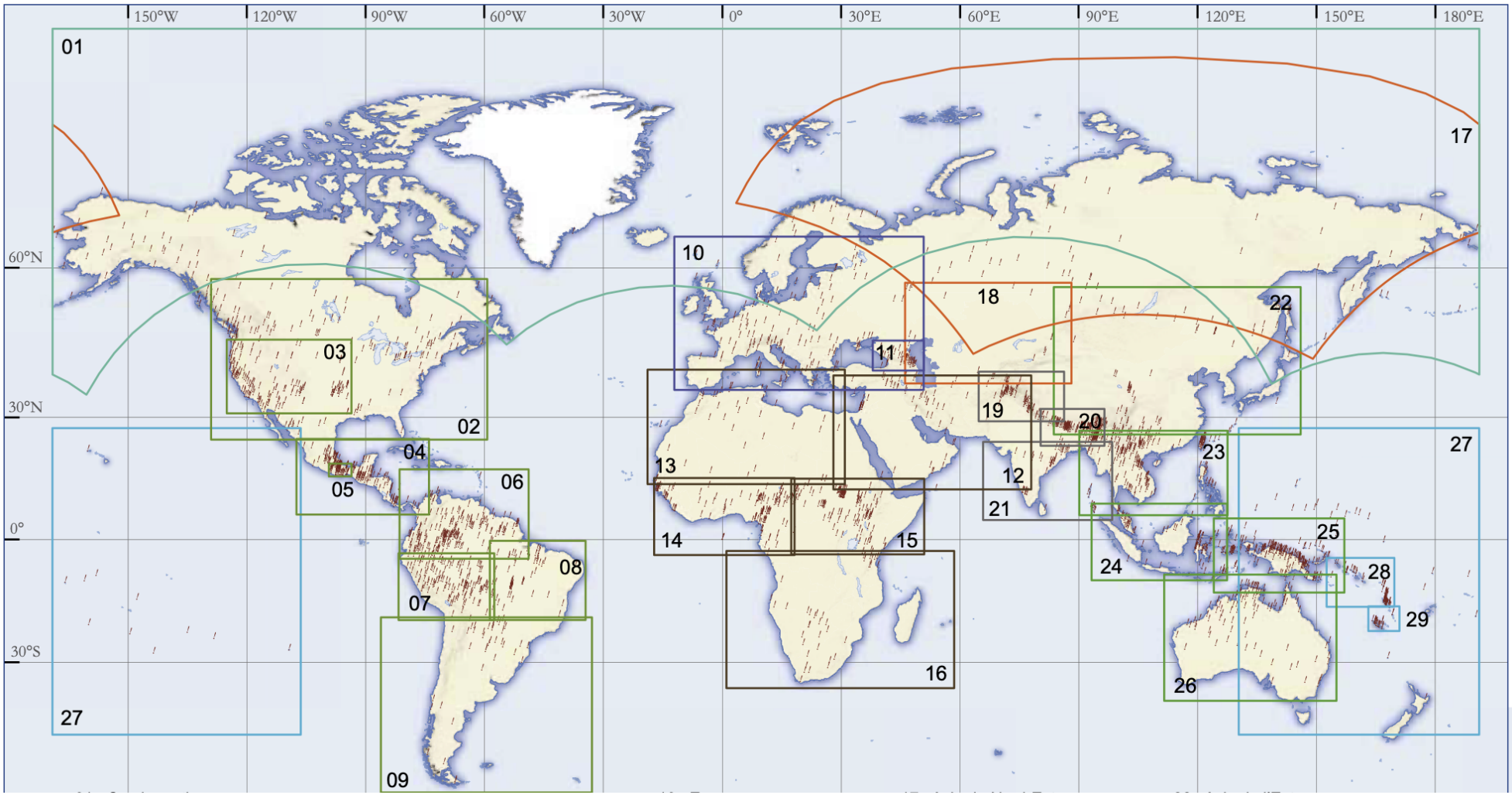
- PhD @ Université Paris-Saclay/CNRS/LISN (FR) (2020–2023)
  Topic: Weakly Supervised Models for Computational Language Documentation

- Post-doc @ TU Munich (DE) (2024–)
  Topic: Parallel sentence mining for (very) low-resource languages

- NLP models and tools for low-resource languages
- **NLP perspective**

Endangered languages

- Around half of the world's languages are endangered



Atlas of the world's languages in danger (UNESCO, 2010)

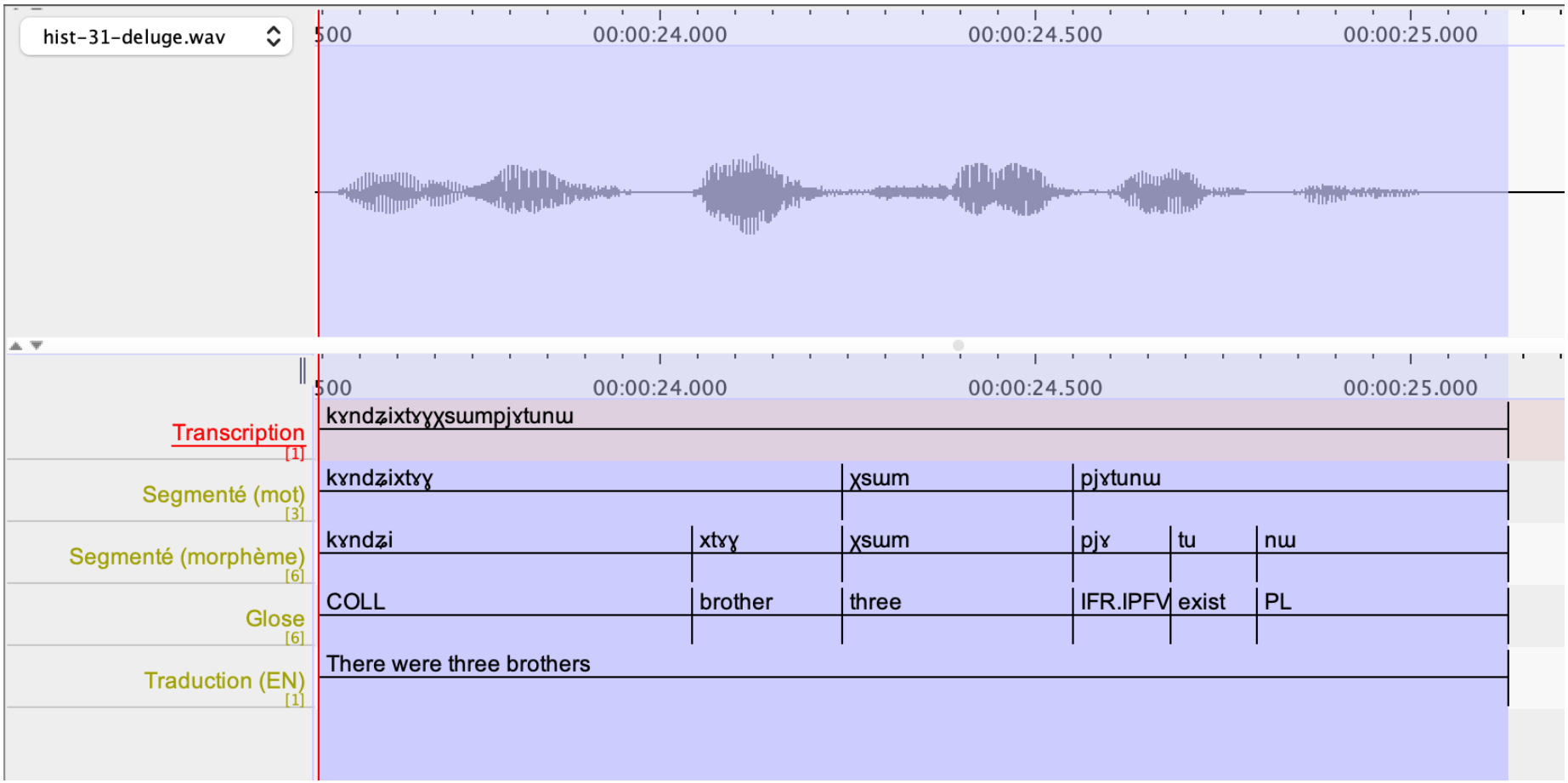
Context for the presentation

- Language Documentation
- Computational Language Documentation (CLD)

- Goal: Reduce annotation bottleneck with partial automation

- Context of the thesis: French-German project (ANR-DFG) 'Computational Language Documentation by 2025' (CLD2025)

Language Documentation



Computational Language Documentation (CLD)

S0	Audio recording	Audio file		
S1	Unsegmented transcription	kɤndʑixtɤɣχsɯmpjɤtunɯ		
S2	Segmented into words	kɤndʑixtɤɣ	χsɯm	pjɤtunɯ
S3	Segmented into words and morphemes	kɤndʑi–xtɤɣ	χsɯm	pjɤ–tu–nɯ
S4	Glossed sentence	COLL–brother	three	IFR.IPFV–exist–PL
S5	Translation (EN)	*There were three brothers.*		

Main annotation tiers of a sentence in (computational) language documentation.

Two tasks:

1. Sequence segmentation (words and morphemes)
2. Automatic gloss generation

Word segmentation

(Okabe et al., 2022)

Word segmentation

S0	Audio recording	Audio file		
S1	Unsegmented transcription	kɤndʑixtɤɣχsɯmpjɤtunɯ		
S2	Segmented into words	kɤndʑixtɤɣ	χsɯm	pjɤtunɯ
S3	Segmented into words and morphemes	kɤndʑi–xtɤɣ	χsɯm	pjɤ–tu–nɯ
S4	Glossed sentence	COLL–brother	three	IFR.IPFV–exist–PL
S5	Translation (EN)	*There were three brothers.*		

- Unsupervised approaches in language documentation

(Godard et al., 2016)

Unsupervised segmentation model: dpseg

dpseg (Goldwater et al., 2009): model based on Dirichlet processes

Sentence representation

i	00	01	02	03	04	05	06	07	08	09	10	1111	12	13	14	15	16	17	18	19	20
c_i	k	ɤ	n	d	ʐ	i	x	t	ɤ	ɣ	χ	s	ɯ	m	p	j	ɤ	t	u	n	ɯ
b_i	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	1

Two hypotheses: presence (1) or absence (0) of a boundary

	…	ɣ	χ	**s**	ɯ	m	…		
presence	…	1	0	**1**	0	1	…	→	χs ɯm
absence	…	1	0	**0**	0	1	…	→	χsɯm

Weak supervision of segmentation models

> *Yet after centuries of colonisation, missionary endeavours, and linguistic fieldwork, all languages have been identified and classified.* ***There is always a wordlist****. We know the language family. Related languages have been studied.* ***There are texts and translations****.* ***There may be linguists and speakers****.*
>
> — *From (Bird, 2020), Decolonising Speech and Language Technology*

- To what extent can we leverage additional resources to improve word segmentation models?

Realistically available resources

1. Boundary annotations:
   - pauses in the recordings (partial boundaries)
   - already segmented sentences (`sentence` supervision)

2. List of words:
   - pre-existing dictionaries
   - words extracted from already segmented sentences (`dictionary` supervision)

Weak supervision strategies

- sentence: all boundaries and non-boundaries are observed for a fraction of sentences

	s	u	b	h	u	l	a	t	s	o	s	a	i	d	i	l	a	m	w	a	n	y	a	a
unsup.	-1	-1	-1	-1	**-1**	-1	**-1**	-1	-1	-1	-1	**-1**	-1	-1	**-1**	-1	**-1**	-1	-1	-1	-1	-1	-1	1
sent.	0	0	0	0	**1**	0	**1**	0	0	0	0	**1**	0	0	**1**	0	**1**	0	0	0	0	0	0	1

- **`dictionary`**: for a supervision dictionary D,
  update the base distribution
  for a **bigram** character model 'learnt' on D

  original unigram $\quad P(\text{'idi'}) \rightarrow P(\text{'i'}) \cdot P(\text{'d'}) \cdot P(\text{'i'})$
  modified bigram $\quad P(\text{'idi'}) \rightarrow P(\text{'id'}) \cdot P(\text{'di'})$

Linguistic material for word segmentation

Two languages currently being documented

- **Mboshi** (Bantu C25) (Godard et al., 2018)
  Tonal Bantu language spoken in the Republic of the Congo
  Example: mosωngωsω nga pora ya nω ye (‘*show me your wound*’)
- **Japhug** (Jacques, 2021)
  Sino-Tibetan language spoken in the western part of China
  Example: kɤndʑi–xtɤɣ χsɯm pjɤ–tu–nɯ

Number of:

- N_{utt}: utterances
- N_{type}: types
- N_{token}: tokens

language	Mboshi	Japhug
N_{utt}	5,130	3,628
N_{type}	5,312	6,739
N_{token}	30,556	28,579

Evaluation metrics for segmentation

Example of a reference and segmented sentence (Mboshi)

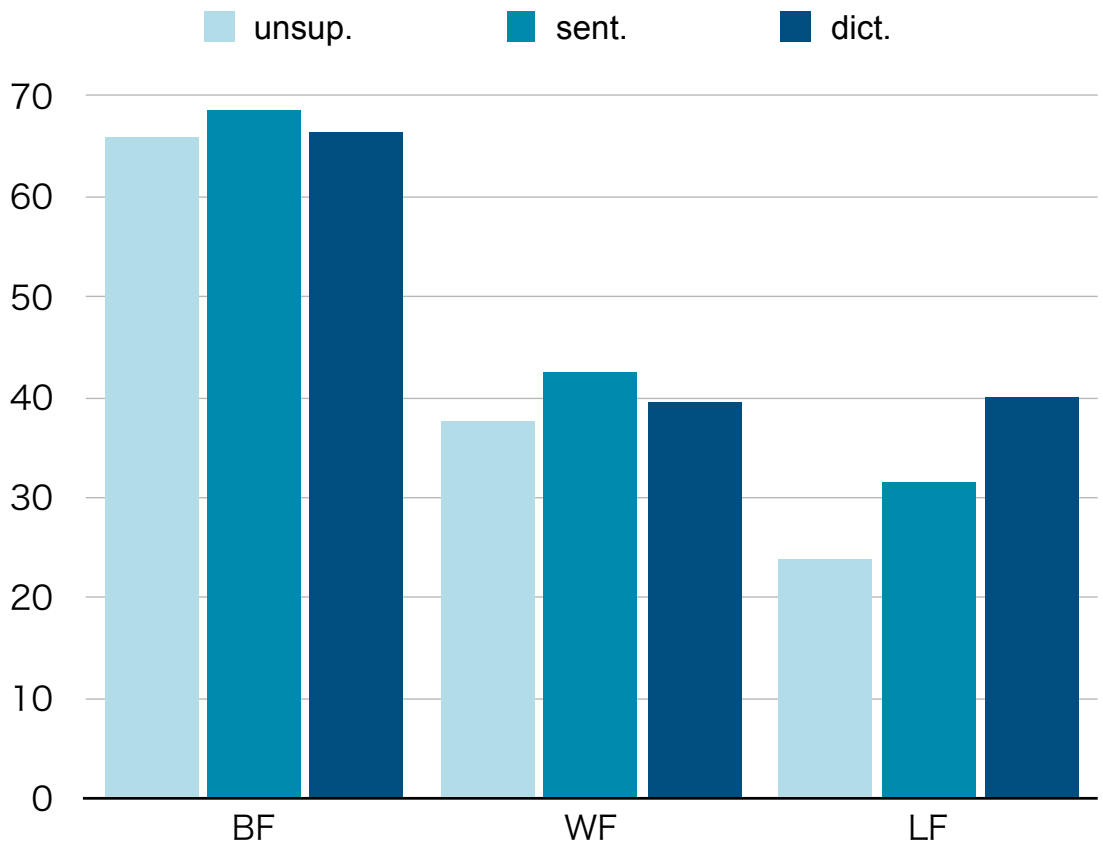
R	Reference	subhu	la	tsosa	idi la	mwanyaa
S	Segmented	sub hu	la	tsosa	idila	mwanyaa

Text statistics & 3 F-scores:

- **BF**: F-score on boundaries
  - R [0, 0, **0**, 0, 1, 0, 1, 0, 0, 0, 0, 1, 0, 0, **1**, 0, 1, 0, 0, 0, 0, 0, 0]
  - S [0, 0, **1**, 0, 1, 0, 1, 0, 0, 0, 0, 1, 0, 0, **0**, 0, 1, 0, 0, 0, 0, 0, 0]
- **WF**: F-score on word tokens
  - R **subhu** la tsosa **idi la** mwanyaa
  - S **sub hu** la tsosa **idila** mwanyaa

  $\Rightarrow$ precision & recall
- **LF**: F-score on word types
  - R {**subhu**, la, tsosa, **idi**, mwanyaa, ...}
  - S {**sub**, **hu**, la, tsosa, **idila**, mwanyaa, ...}

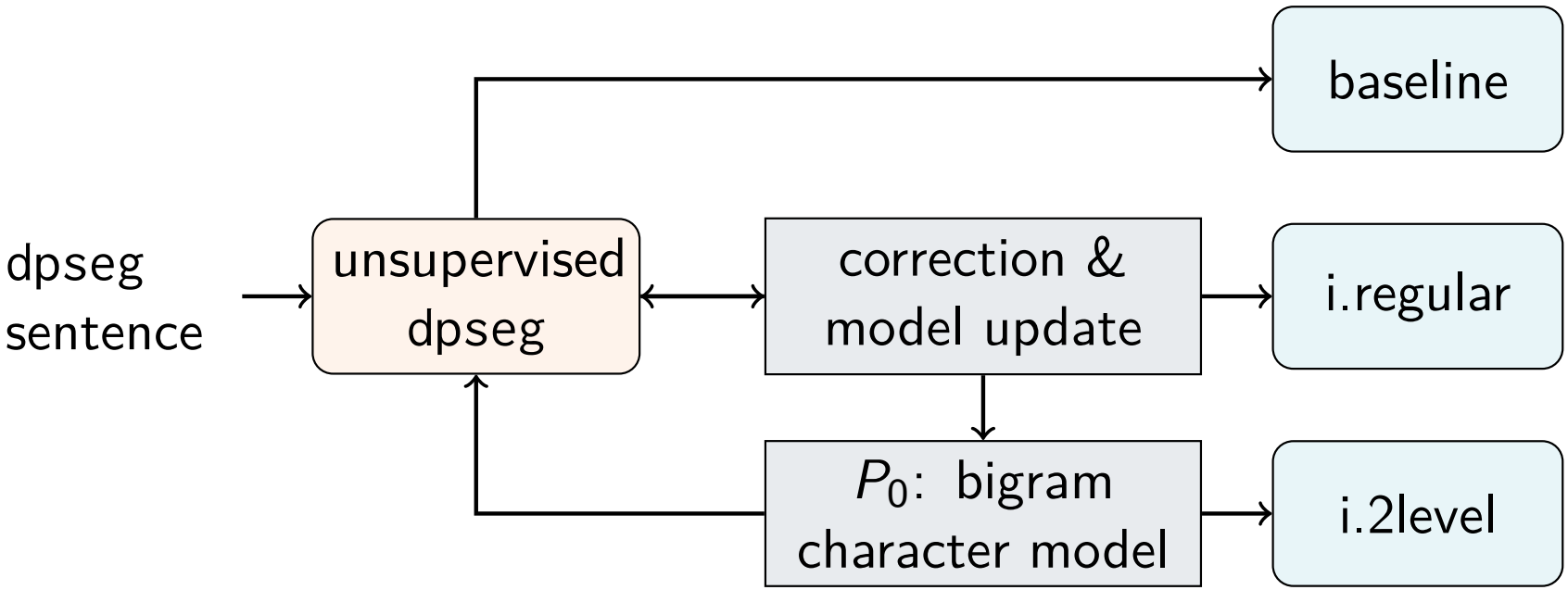
Weak supervision results (Mboshi)



super.	unsup.	sent.	dict.
WL[1]	3.74	3.78	5.11
TL[1]	4.61	4.87	6.60
N_{type}	1,980	2,237	4,620
N_{token}	34.2k	33.8k	25.0k

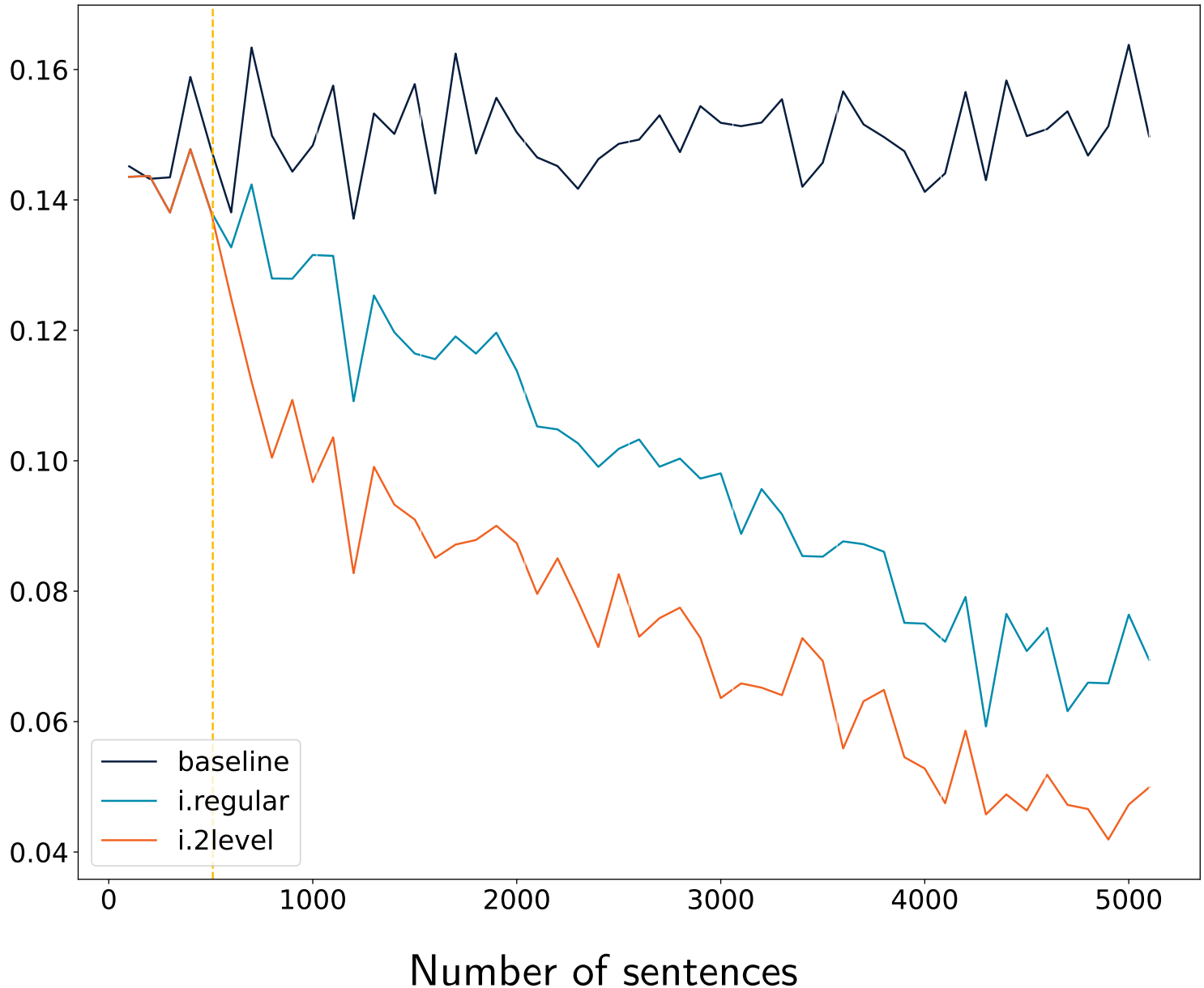
[1] WL: average token length; TL: average type length

Incremental learning



- Error rate: $\frac{\text{number of errors over 100 sentences}}{\text{length of 100 sentences}}$

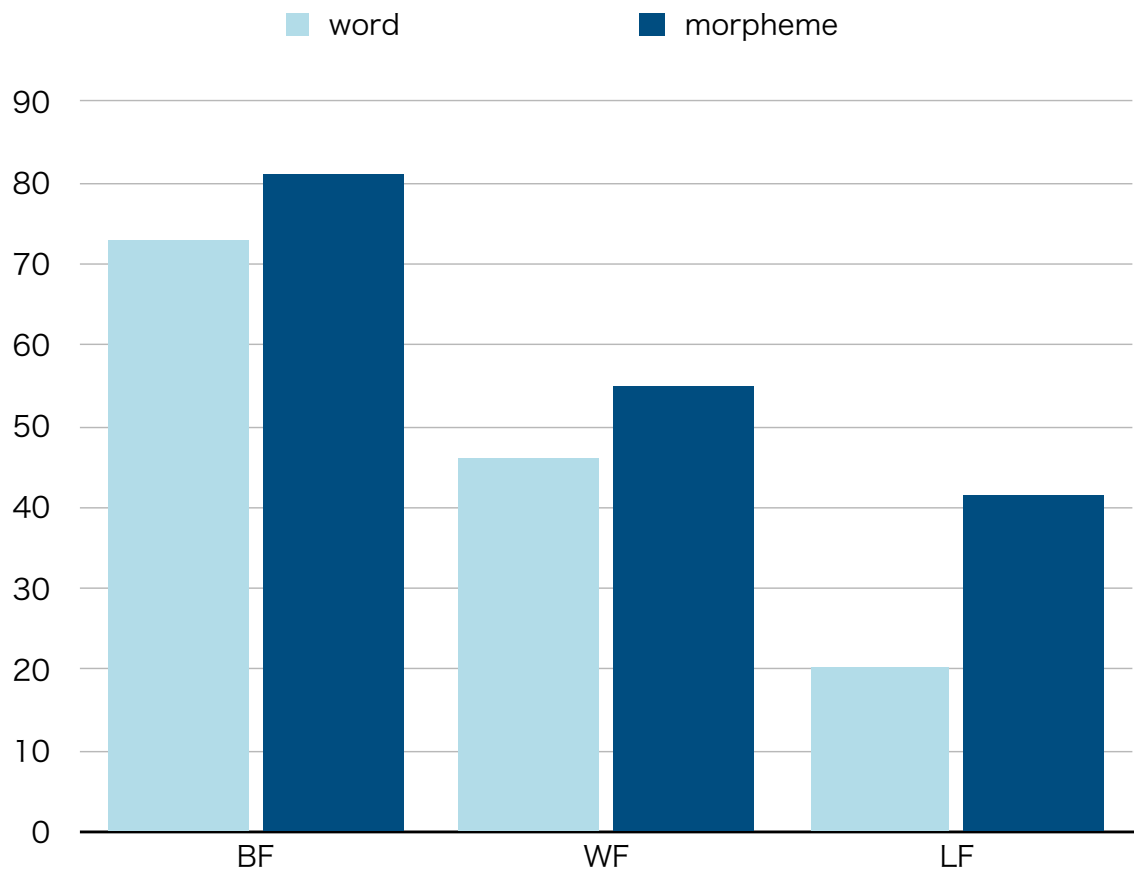
Incremental learning: Results (Mboshi)



Segmented units: words or morphemes?

- Reference Japhug text, segmented in words or morphemes

	reference (word)	reference (morpheme)	dpseg
WL	4.60	2.85	3.34



Segmenting into words and morphemes

(Okabe and Yvon, 2023a)

Two-level segmentation objective

S0	Audio recording	Audio file		
S1	Unsegmented transcription	kɤndʑixtɤɣχsɯmpjɤtunɯ		
S2	Segmented into words	kɤndʑixtɤɣ	χsɯm	pjɤtunɯ
S3	Segmented into words and morphemes	kɤndʑi–xtɤɣ	χsɯm	pjɤ–tu–nɯ
S4	Glossed sentence	COLL–brother	three	IFR.IPFV–exist–PL
S5	Translation (EN)	*There were three brothers.*		

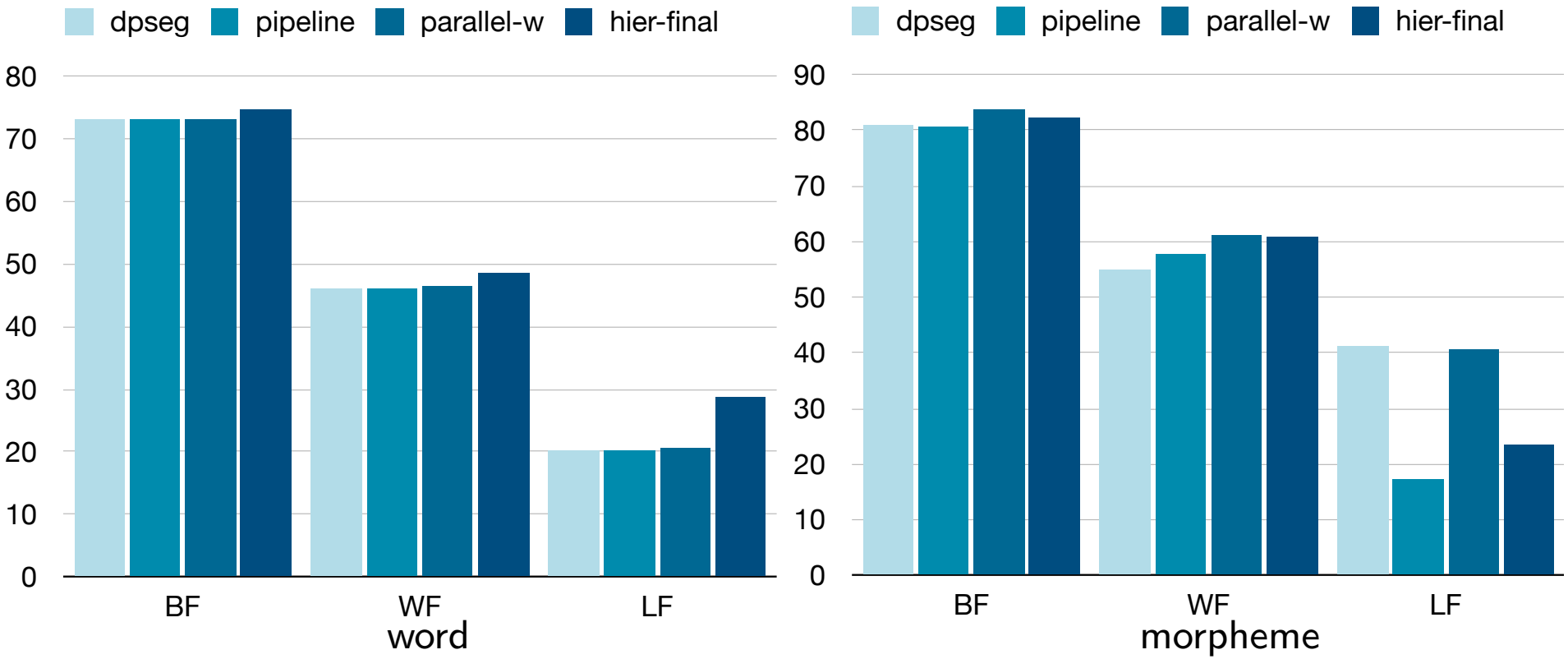
→ Can we better differentiate the two levels of segmentation?

→ Can we better segment into words?

Two-level segmentation models

- **Pipeline** system
  1. segmenting first into words
  2. then segmenting found **types** into morpheme
- **Parallel** approach
  - one model per level
  - merging consistency through rules
- **Hierarchical** model
  - integrating the relationship between words and morphemes
  - words are made of one or more morphemes

Unsupervised results (Japhug)



model	reference		parallel-w		hier-final	
level	word	morph.	word	morph.	word	morph.
WL	4.73	2.90	3.34	2.98	3.93	2.32

Conclusion - sequence segmentation

- Improving segmentation models through weak supervision using realistic additional resources such as word lists
- Incremental learning to simulate correction by a linguist
- Better word-level segmentation by incorporating the relationship between words and morphemes
- Difficulty to distinguish between words and morphemes from statistical cues only

Automatic gloss generation

(Okabe and Yvon, 2023b)

Automatic gloss generation

S0	Audio recording	Audio file		
S1	Unsegmented transcription	kɤndʑixtɤɣχsɯmpjɤtunɯ		
S2	Segmented into words	kɤndʑixtɤɣ	χsɯm	pjɤtunɯ
S3	Segmented into words and morphemes	kɤndʑi–xtɤɣ	χsɯm	pjɤ–tu–nɯ
S4	Glossed sentence	COLL–brother	three	IFR.IPFV–exist–PL
S5	Translation (EN)	*There were three brothers.*		

- Interlinear glosses are costly to obtain

First task: binary classification

Input **S3**	Segmented sentence	nesi–s	ɬˤono	uži	zow–n
Output **S4'**	Gloss type	LEX–GRAM	LEX	LEX	LEX–GRAM

- Label set: $\mathcal{Y} = \{\text{LEX}, \text{GRAM}\}$
- One-to-one correspondence between morphemes and labels
- Classic sequence labelling model: Conditional Random Field (CRF)

Second task: with grammatical glosses

Input **S3**	Segmented sentence	nesi–s	ɬˤono	uži	zow–n
Output **S4"**	Gram. gloss + LEX	LEX–GEN1	LEX	LEX	LEX–PST.UNW

- Label set: $\mathcal{Y} = \{\text{LEX}\} \bigcup \mathcal{Y}_G$
  where $\mathcal{Y}_G$ is the set of all grammatical glosses[2]

- Methodology in (Moeller and Hulden, 2018) or (Barriga Martínez et al., 2021)

[2] More than 100 labels in our Tsez data.

Predicting lexical glosses

Input 1	**S3**	Segmented sentence	nesi–s	ɬˢono	uži	zow–n
Input 2	**S5**	Translation (EN)	*He had three sons.*			
Output	**S4**	Glossed sentence	he.OBL–GEN1	three	son	be.NPRS–PST.UNW

⇒ The label set cannot be fixed beforehand:
the variety of lexical glosses is almost unbounded

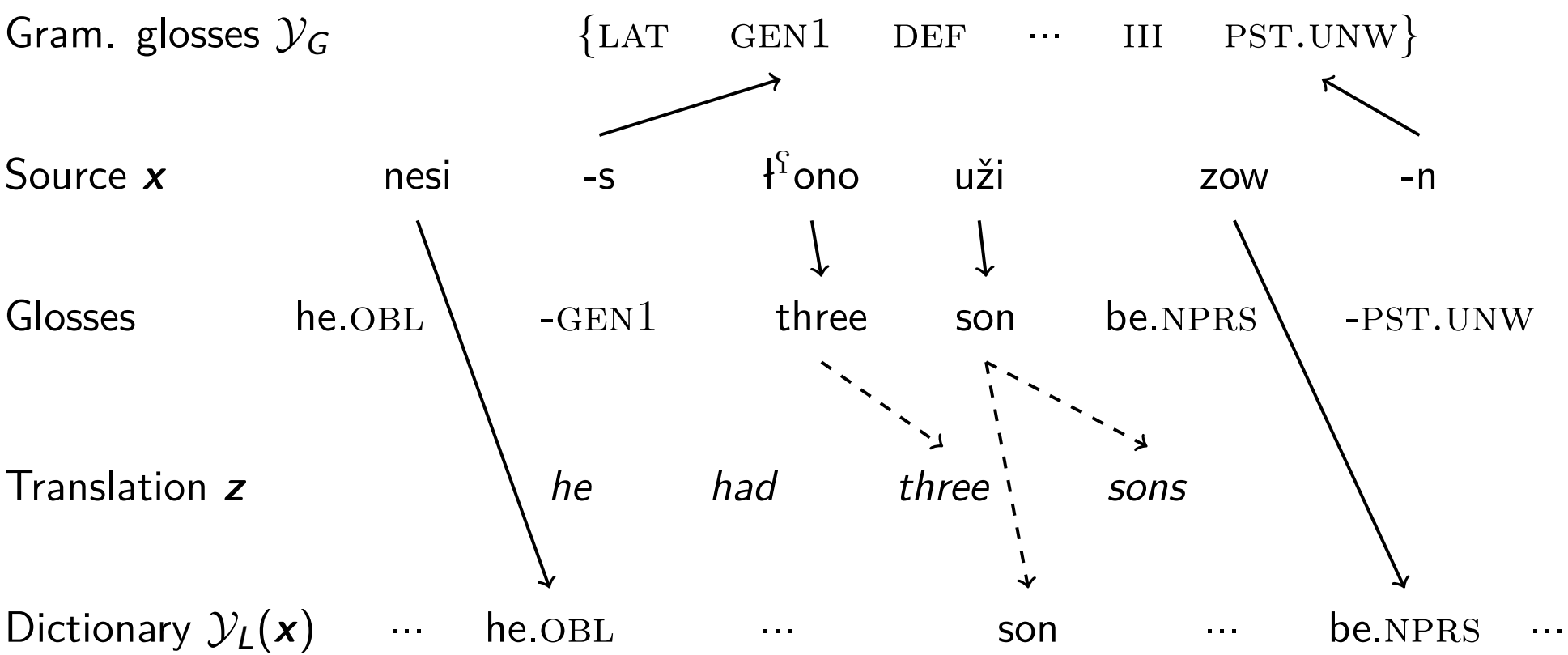
Hypothesis:

- assume that lexical glosses may be obtained from the translation

(McMillan-Major, 2020) and (Zhao et al., 2020)

Illustration of our approach

Based on a variant of CRFs (Lavergne et al., 2013) → **dynamic** label set



Feature functions in the model

input	selected input features					output
m	*t*	*l*	*d*	*e*	*ps*	*g*
nesi	0	4	nes	esi	1/4	he.OBL
s	1	1	s	s	1/4	GEN1
ɬˠono	F	5	ɬˠo	ono	2/4	three
uži	F	3	uži	uži	2/4	son
zow	0	3	zow	zow	3/4	be.NPRS
n	1	1	n	n	4/4	PST.UNW

Constructing the set of possible labels (search space)

- $\mathcal{Y}_G$: **grammatical glosses**, a shared (finite) set for all sentences
- $\boldsymbol{z}$: lemmas of the words in the **translation**
- $\mathcal{Y}_L(\boldsymbol{x})$: the most frequent lexical glosses associated with source morphemes seen in training sentences (**dictionary**)

Three versions:

- **S1**: dictionary ($\mathcal{Y}_L(\boldsymbol{x})$) only
- **S2**: translation ($\boldsymbol{z}$) only
- **S3**: both sources

Languages from the 2023 SIGMORPHON Shared Task

5 among 7 languages of the 2023 SIGMORPHON Shared Task:
Tsez (ddo), Gitksan (git), Lezgi (lez), Natugu (ntu), and Uspanteko (usp)

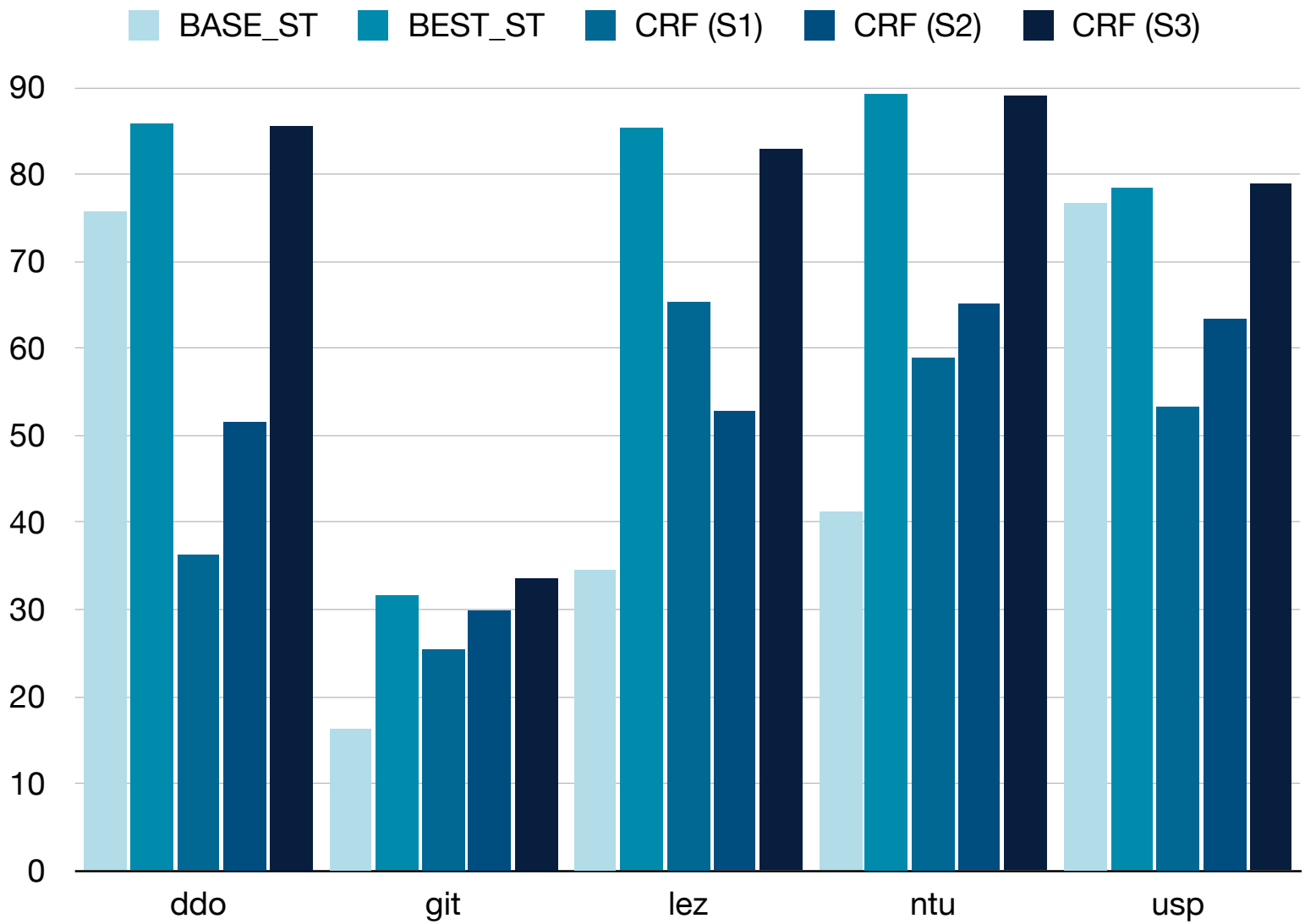
language	ddo	git	lez	ntu	usp
train	3,558	31	701	791	9,774
development	445	42	88	99	232
test	445	37	87	99	633
target language	EN	EN	EN	EN	ES

Number of sentences and translation language for each language.

- Size of training data ranging from 31 to 9,774 sentences
- **Two** (target) documentation languages

Main glossing experimental results

Evaluation: word-level accuracy scores



Glossing in practice: NLP gap

> '***Documentary linguistics will not benefit from advances in NLP*** *until significant investments are made in* ***developing application software*** *which can compete with existing apps in functionality and* ***provide first-class support for NLP model integration***.'

— *From (Gessler, 2022), Closing the NLP Gap: Documentary Linguistics and NLP Need a Shared Software Infrastructure*

→ Lowering the technical barrier for glossing

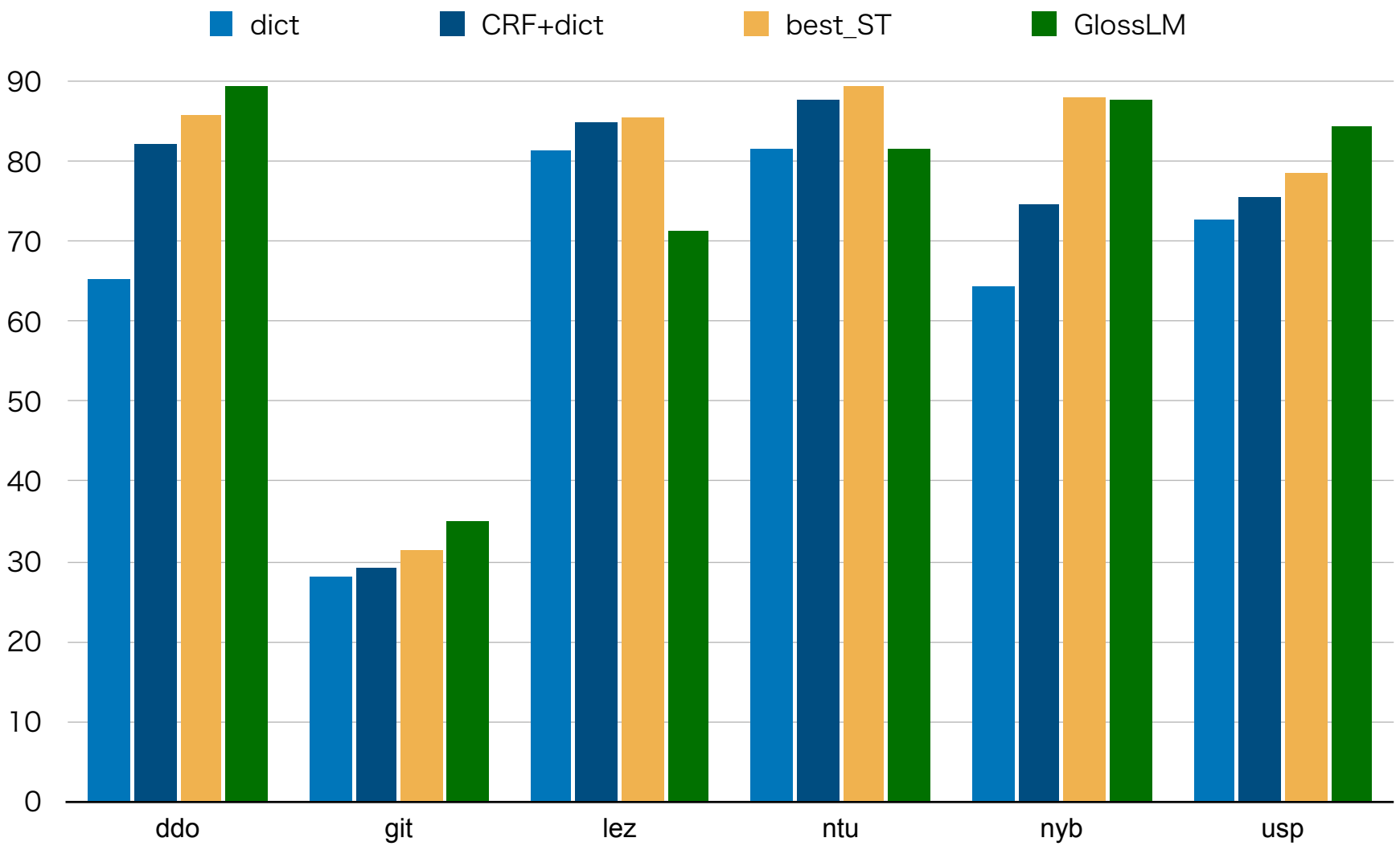
(Asadpour et al., 2025)

Second task: with grammatical glosses (reminder)

Input **S3**	Segmented sentence	nesi–s	ɬˤono	uži	zow–n
Output **S4"**	Gram. gloss + LEX	LEX–GEN1	LEX	LEX	LEX–PST.UNW

- Label set: $\mathcal{Y} = \{\text{LEX}\} \bigcup \mathcal{Y}_G$
  where $\mathcal{Y}_G$ is the set of all grammatical glosses
- Methodology in (Moeller and Hulden, 2018) or (Barriga Martínez et al., 2021)
- **Using a classic sequence labelling model: CRF + dictionary for lexical labels**

Compared to the state-of-the-art (6 languages)



Interpretability: features and transitions

- For a documented language: Mukrī Kurdish (Central Kurdish dialect)
- Interpreting learnt features and label transitions

Feature		**Transition**	
source feature	gloss	gloss_1	gloss_2
morph: m	1SG	EZ	REFL
morph: ew	DEM	IND	–
morph: emin	1SG	INDF.PRO	INDF.PRO
morph: eto	2SG	PVB	–
morph: de	IND	=	3SG

Conclusion - automatic gloss generation

- Restriction of the possible label set *locally* (search space) thanks to a model based on CRFs
- Competitive scores with respect to the best results in the 2023 SIGMORPHON Shared Task on glossing
- Multilingual pre-training of the model with external glossed data
- Trade-off: Simple implementation based on a CRF and a dictionary for glossing

More on interlinear glosses

Interlinear glosses in NLP

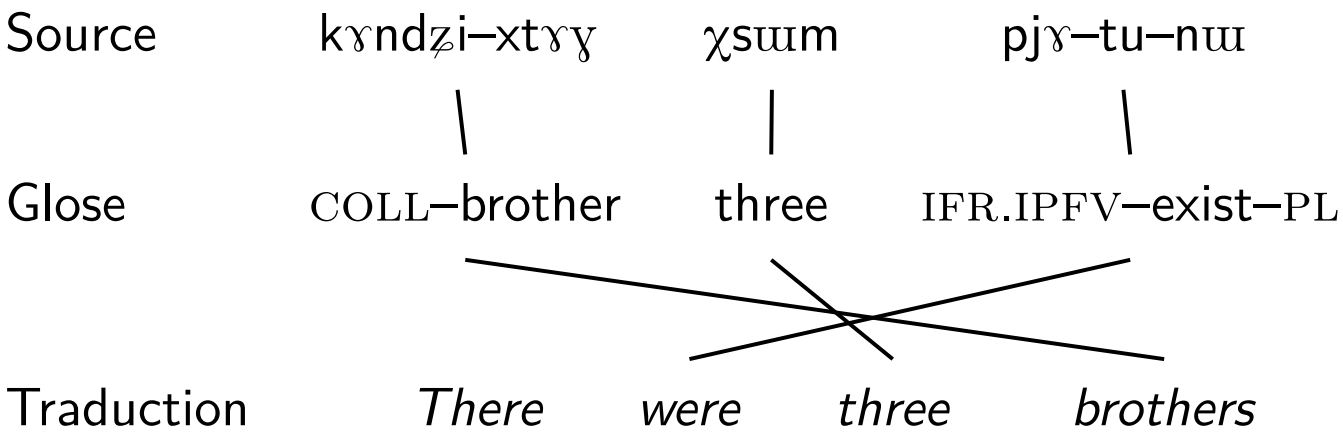
- Increased interest in interlinear glosses since 2023 (Shared Task)

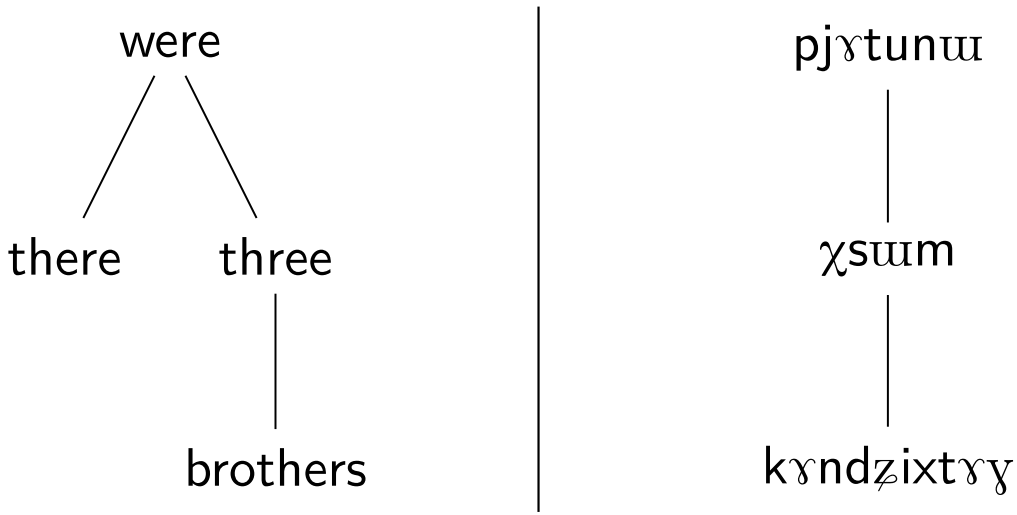
Selected works:

- 2024:
  - Glossing from speech data: **wav2gloss** (He et al., 2024)
  - State-of-the-art glossing model: **GlossLM** (Ginn et al., 2024a)
  - Using LLMs for glossing (Ginn et al., 2024b)
- 2025:
  - Position paper: (Rice et al., 2025)
    '*Despite strong performance on standard metrics, GlossLM falls short for real-world use*'

Keynote presentation at **Field Matters 2026**
Automated glossing for language documentation: Historical perspective, state of the art, and the importance of the user – Alexis Palmer

Projection of annotations through glosses





e.g., methodology from (Xia and Lewis, 2007)

IGT to help Machine Translation (MT)

Source	kɤndʑi–xtɤɣ	χsɯm	pjɤ–tu–nɯ
Glosses	COLL–brother	three	IFR.IPFV–exist–PL
Translation	*There were three brothers.*		

- Glosses can be used as a bridge between two languages
- Pivoting with a pipeline approach (neural MT)

  (Zhao et al., 2020), (Özer et al., SIGUL 2026)
- Generating (or relying on generated) glosses for MT: GrammaMT (Ramos et al., 2025)

Conclusion

- Two Computational Language Documentation tasks:
  word (and morpheme) **segmentation** and **automatic gloss generation**
- Leveraging realistically available resources through a **weak supervision** of segmentation models
- **Joint segmentation** into words and morphemes to better differentiate the nature of segmented units
- Automatic gloss generation using **feature-based models** with competitive results
- Glosses as a **bridge** between the source and target languages in NLP applications

Perspectives

- Using glosses as pivot between the source and target languages
- Collaboration with linguists and other stakeholders for low-resourced languages
- Integrating the tools in actual annotation platforms used by linguists

Going beyond: Field linguistics and NLP

Field Matters workshop

Thank you!