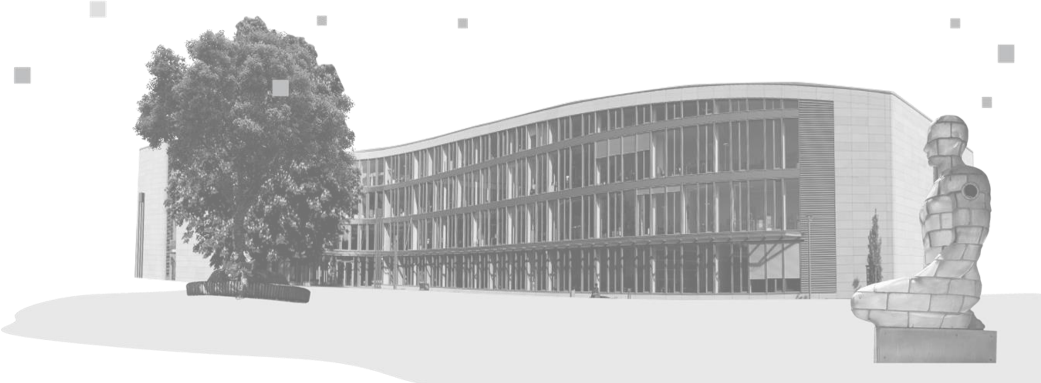
GUIDE: Creating Semantic Domain Dictionaries for Low-Resource Languages

Jonathan Janetzki, Gerard de Melo, Joshua Nemecek, Daniel Whitenack



Design IT.
Create Knowledge.

Joshua Nemecek

Gerard de Melo

Jonathan Janetzki

Daniel Whitenack

Agenda

1. The Global Language Documentation Gap
2. Dataset Characteristics
3. Graph Building
4. Dictionary Entry Creation
5. Evaluation
6. Conclusion

”Languages shape our tools,

and our tools shape languages.”

”Languages shape our tools,

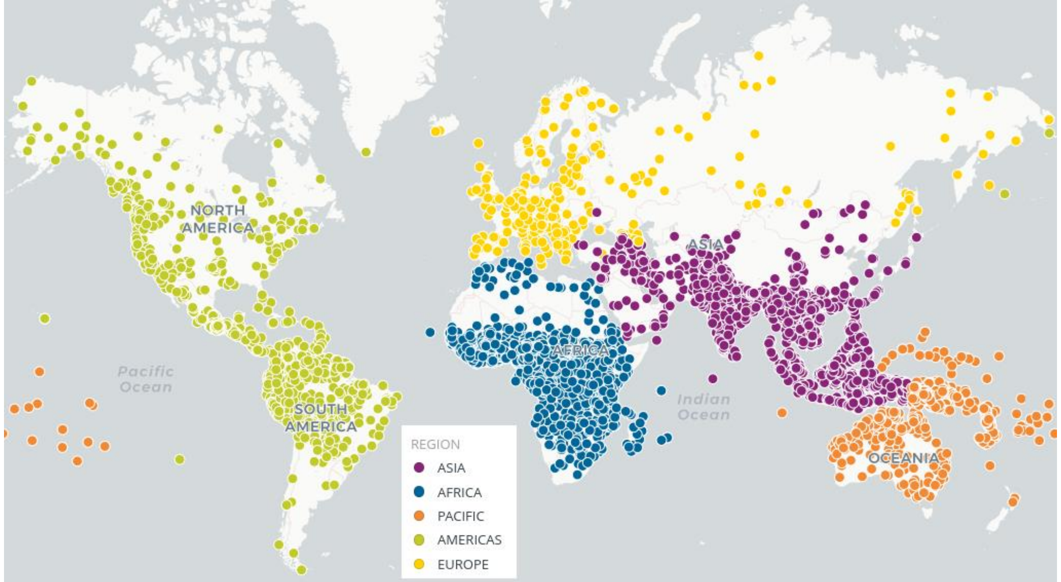
and our tools shape languages.”

— ChatGPT

GUIDE: Graph-based Unified Indigenous Dictionary Engine

The Global Language Documentation Gap

All Living Languages

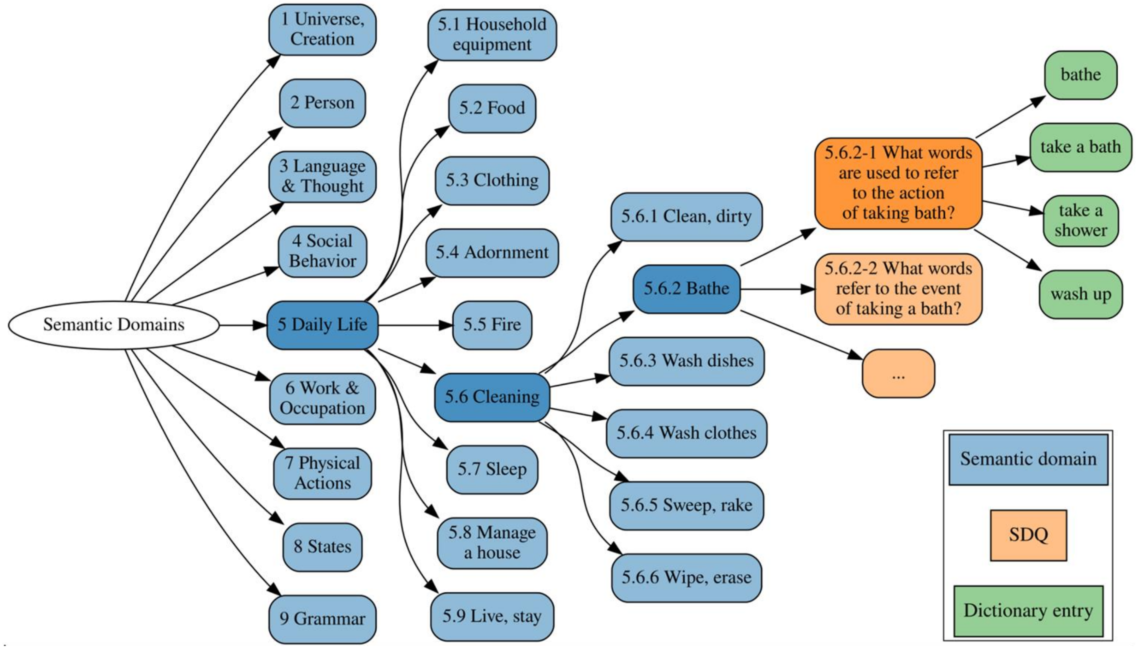


https://www.ethnologue.com/insights/how-many-languages/

- There are 7,168 languages on Earth.

- **> 7,000 low-resource** languages

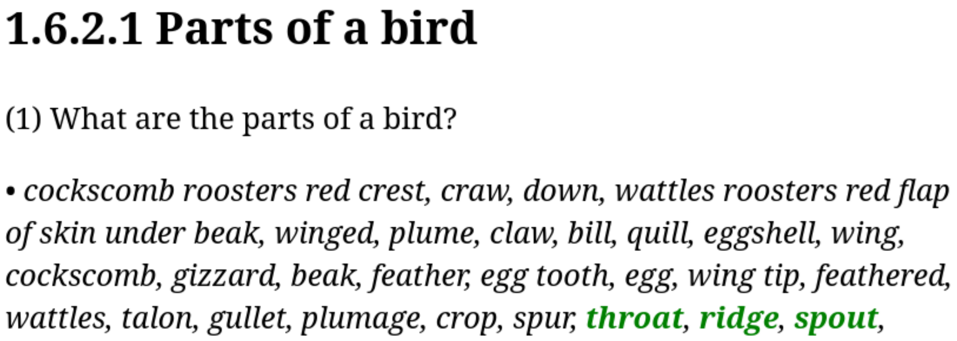
SIL’s Semantic Domains



based on [Moe10]

- Semantic domains are a **tree-structured** ontology.
- *“word-SDQ link”* = dictionary entry
- SDQs allow building highly **multiparallel** dictionaries.
- Words often have **no 1:1 translations** but have different semantic ranges.

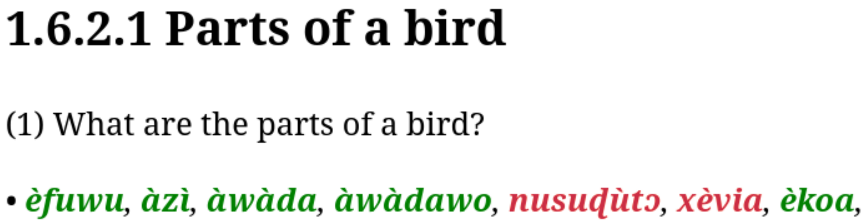
Main Contributions



1.6.2.1 Parts of a bird

(1) What are the parts of a bird?

- *cockscomb roosters red crest, craw, down, wattles roosters red flap of skin under beak, winged, plume, claw, bill, quill, eggshell, wing, cockscomb, gizzard, beak, feather, egg tooth, egg, wing tip, feathered, wattles, talon, gullet, plumage, crop, spur,* ***throat*****,** ***ridge*****,** ***spout*****,**



1.6.2.1 Parts of a bird

(1) What are the parts of a bird?

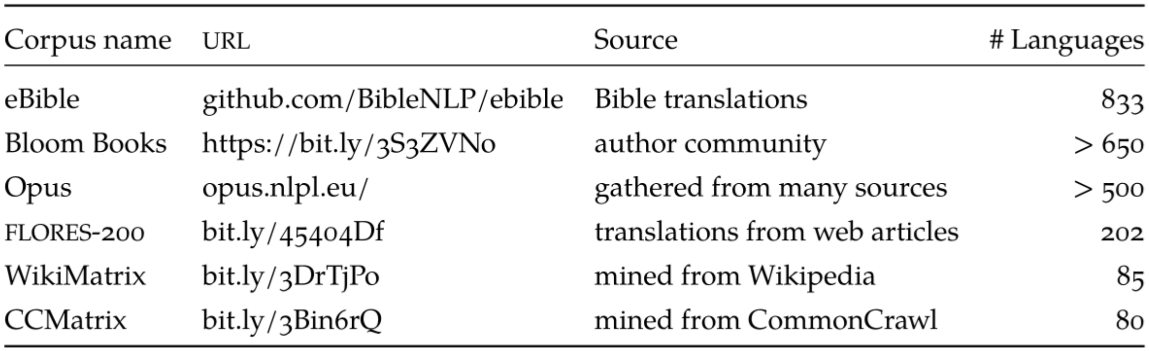
- ***èfuwu*****,** ***àzì*****,** ***àwàda*****,** ***àwàdawo*****,** ***nusuḍùtɔ*****,** ***xèvia*****,** ***èkoa*****,**

(feathers, egg, wing, wings, greedy, bird, gizzard)

- GUIDE finds **missing word-SDQ links** with an avg. precision of 0.68.
- GUIDE finds word-SDQ links **in unseen languages** with an avg. precision of 0.60.
- GUIDE predicts **33,000 correct and new word-SDQ links** in 20 languages.
- We establish a **new benchmark**.
- Open-source: https://github.com/janetzki/guide

Dataset Characteristics

Dataset Review

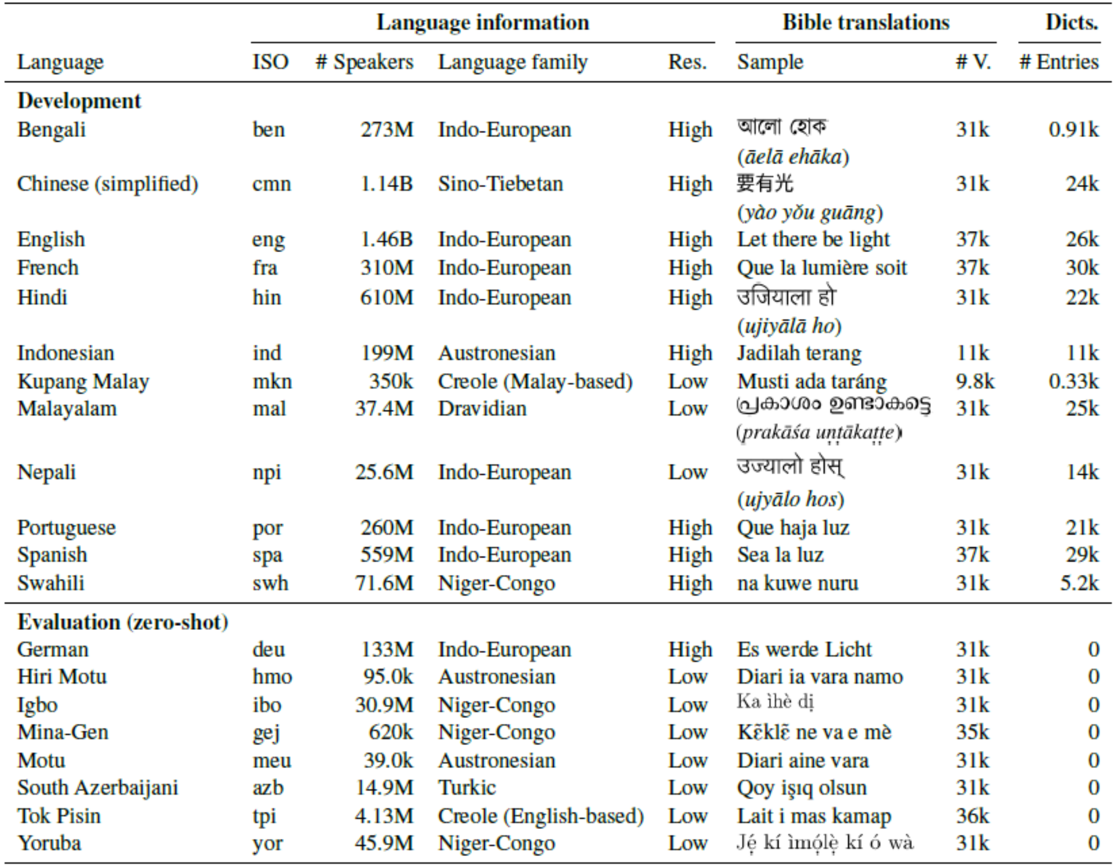


Corpus name	URL	Source	# Languages
eBible	github.com/BibleNLP/ebible	Bible translations	833
Bloom Books	https://bit.ly/3S3ZVNo	author community	> 650
Opus	opus.nlpl.eu/	gathered from many sources	> 500
FLORES-200	bit.ly/45404Df	translations from web articles	202
WikiMatrix	bit.ly/3DrTjPo	mined from Wikipedia	85
CCMatrix	bit.ly/3Bin6rQ	mined from CommonCrawl	80

adapted from [Haddow22]

- We chose the **eBible corpus**.
- **Other corpora** cover fewer languages.

Dataset Size

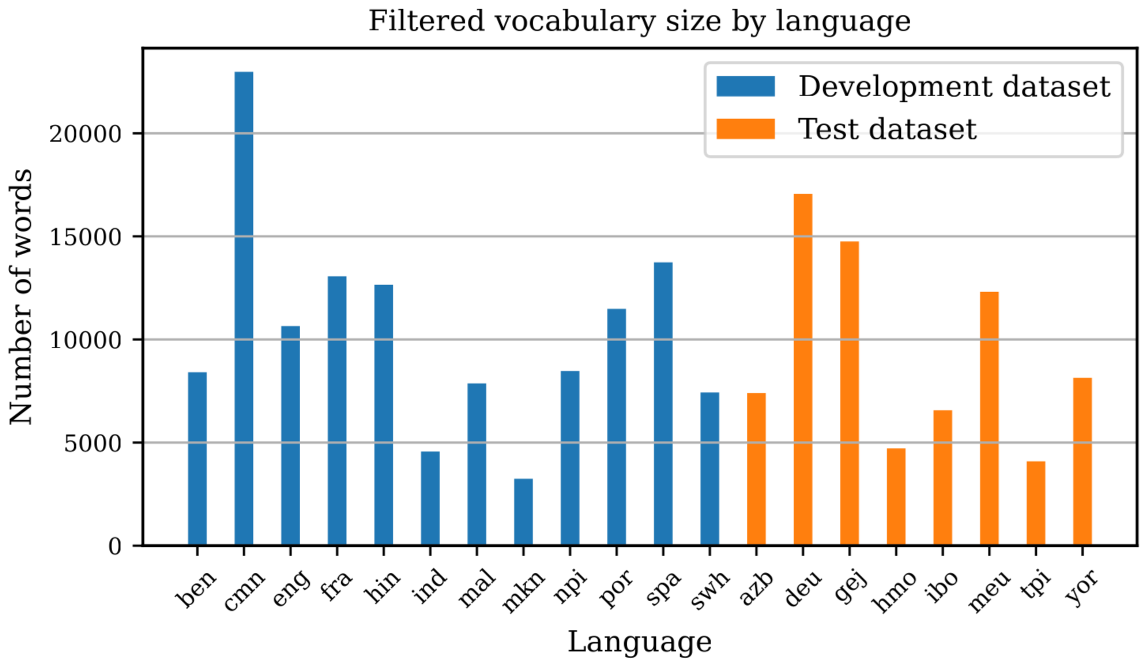


	Language information				Bible translations		Dicts.
Language	ISO	# Speakers	Language family	Res.	Sample	# V.	# Entries
Development							
Bengali	ben	273M	Indo-European	High	আলো হোক (*āelā ehāka*)	31k	0.91k
Chinese (simplified)	cmn	1.14B	Sino-Tiebetan	High	要有光 (*yào yŏu guāng*)	31k	24k
English	eng	1.46B	Indo-European	High	Let there be light	37k	26k
French	fra	310M	Indo-European	High	Que la lumière soit	37k	30k
Hindi	hin	610M	Indo-European	High	उजियाला हो (*ujiyālā ho*)	31k	22k
Indonesian	ind	199M	Austronesian	High	Jadilah terang	11k	11k
Kupang Malay	mkn	350k	Creole (Malay-based)	Low	Musti ada taráng	9.8k	0.33k
Malayalam	mal	37.4M	Dravidian	Low	പ്രകാശം ഉണ്ടാകട്ടെ (*prakāśa uṇṭākaṭṭe*)	31k	25k
Nepali	npi	25.6M	Indo-European	Low	उज्यालो होस् (*ujyālo hos*)	31k	14k
Portuguese	por	260M	Indo-European	High	Que haja luz	31k	21k
Spanish	spa	559M	Indo-European	High	Sea la luz	37k	29k
Swahili	swh	71.6M	Niger-Congo	High	na kuwe nuru	31k	5.2k
Evaluation (zero-shot)							
German	deu	133M	Indo-European	High	Es werde Licht	31k	0
Hiri Motu	hmo	95.0k	Austronesian	Low	Diari ia vara namo	31k	0
Igbo	ibo	30.9M	Niger-Congo	Low	Ka ìhè dị	31k	0
Mina-Gen	gej	620k	Niger-Congo	Low	Kɛ̃klɛ̃ ne va e mè	35k	0
Motu	meu	39.0k	Austronesian	Low	Diari aine vara	31k	0
South Azerbaijani	azb	14.9M	Turkic	Low	Qoy işıq olsun	31k	0
Tok Pisin	tpi	4.13M	Creole (English-based)	Low	Lait i mas kamap	36k	0
Yoruba	yor	45.9M	Niger-Congo	Low	Jẹ́ kí ìmọ́lẹ̀ kí ó wà	31k	0

- 12 development languages
- 8 zero-shot evaluation languages
- 10 low-resource languages
- 7 language families

Language Information based on [Eberhard23]

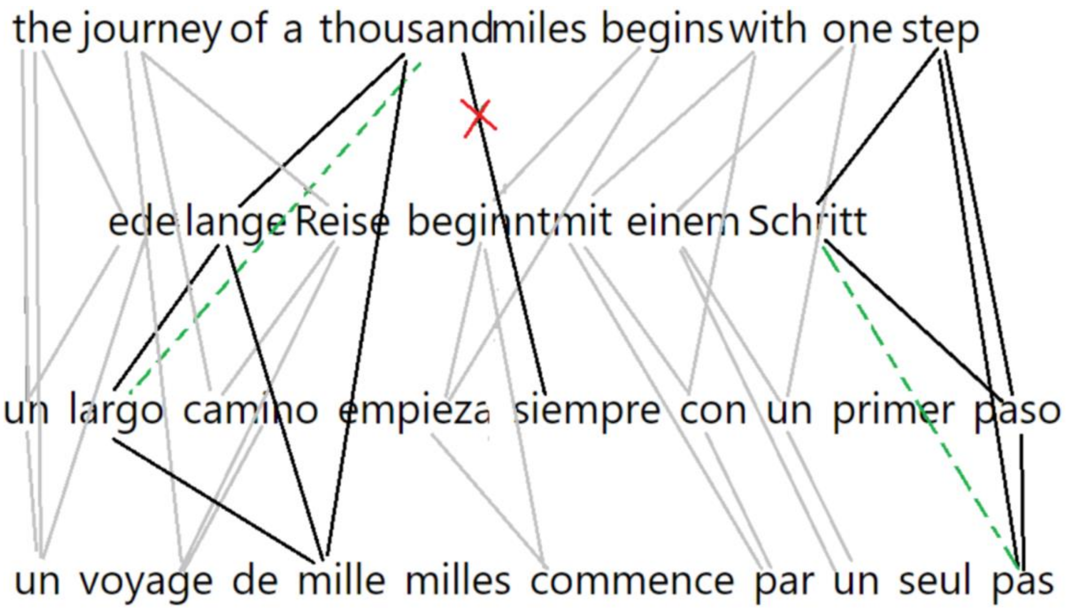
Language Distribution



- **~ 200,000** words
- **~10,000 words** per language on average

Graph Building

Eflomal Word Aligner

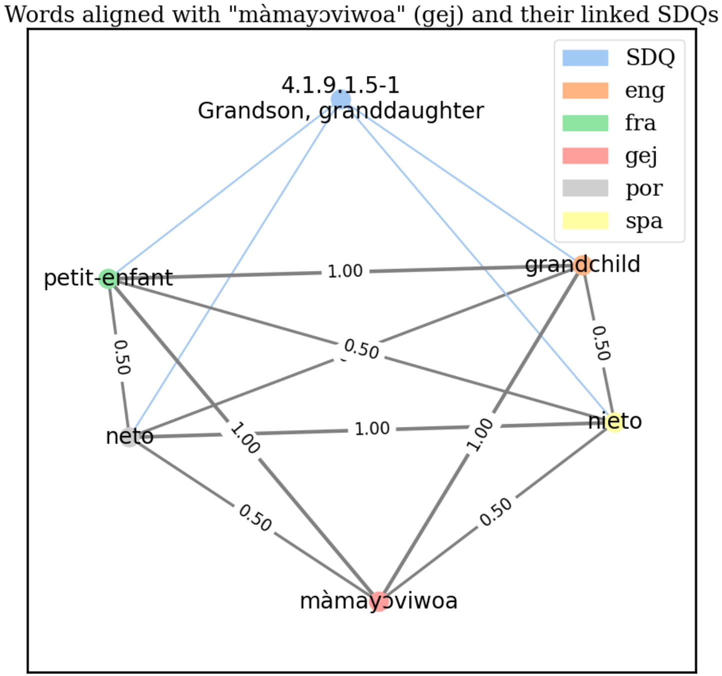


the journey of a thousand miles begins with one step
ede lange Reise beginnt mit einem Schritt
un largo camino empieza siempre con un primer paso
un voyage de mille milles commence par un seul pas

taken from [Imani21]

- Dotted lines = missing alignments
- ❌ = incorrect alignment

Graph Structure



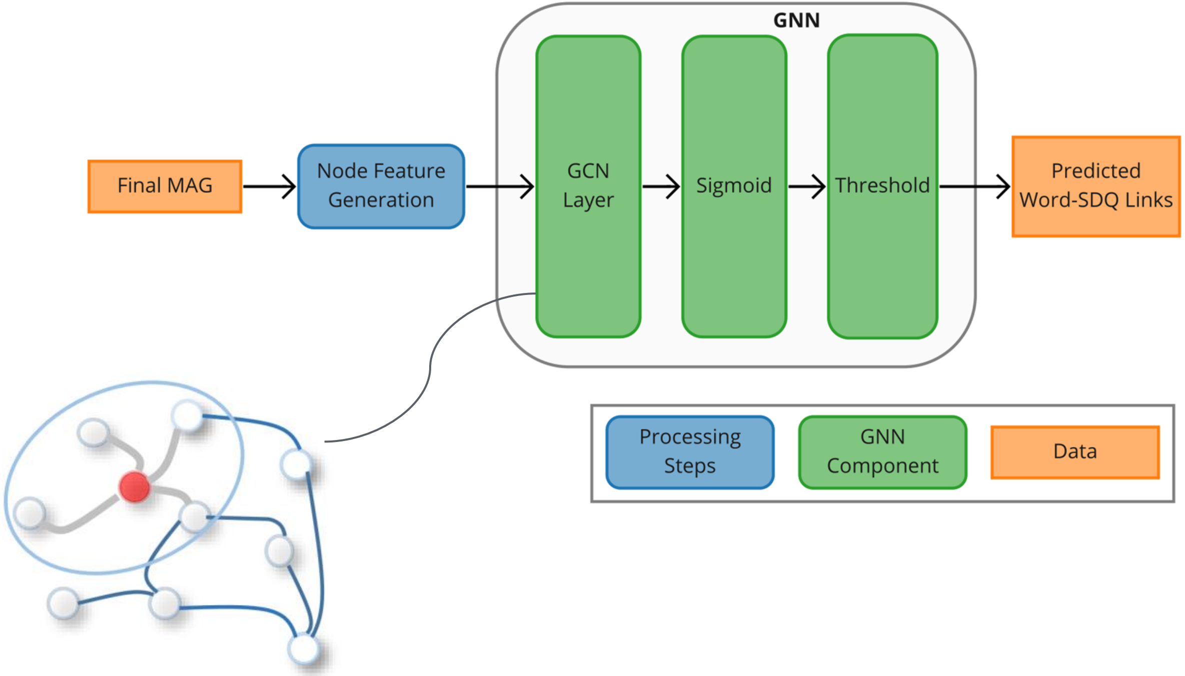
- GUIDE creates a Multilingual Alignment Graph (MAG).
- 1 node = 1 word
- 1 gray edge = n alignments
- edge weight = normalized weight

Just for readability:

- blue node = SDQ (“*What words refer to the children of your children?”*)
- blue edges = word-SDQ links

Dictionary Entry Creation

Model Architecture

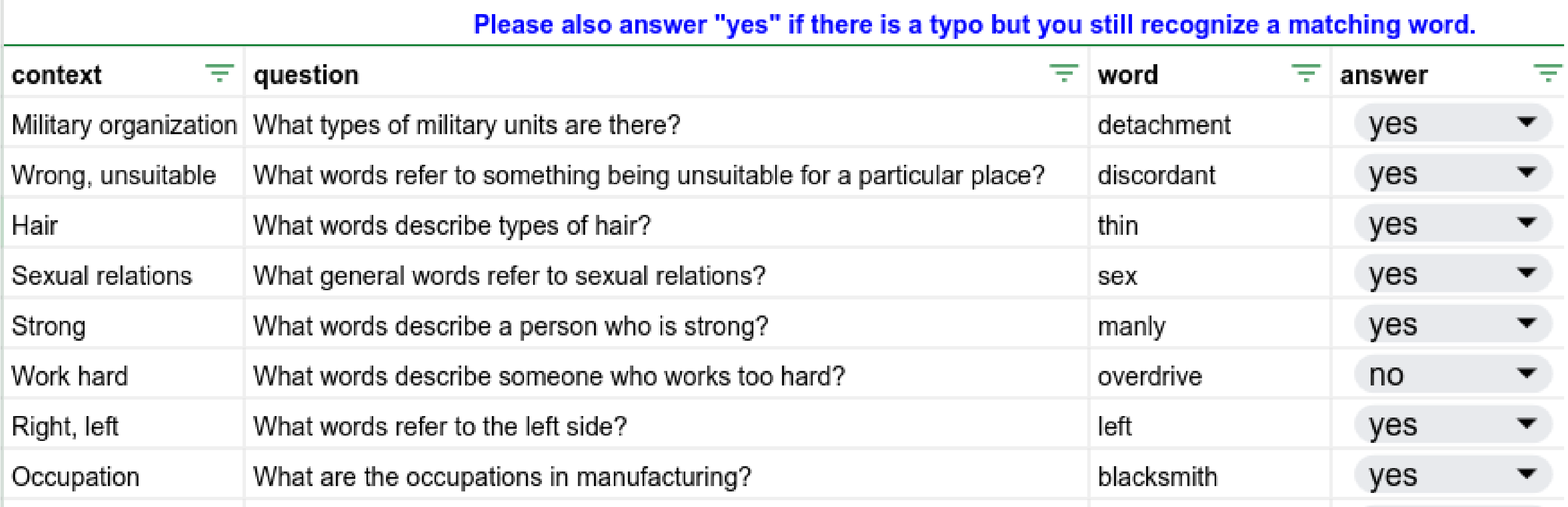


taken from [Wu21]

- GUIDE predicts word-SDQ links using a **single-layer GCN**.
- Threshold = 0.999
- Four node features:
  - Node degree
  - Weighted node degree (i.e., sum of adjacent weights)
  - SDQ count
  - SDQ links
- 7,428 × 7,425 parameters in the weight matrix

Evaluation

Manual Evaluation with Questionnaires

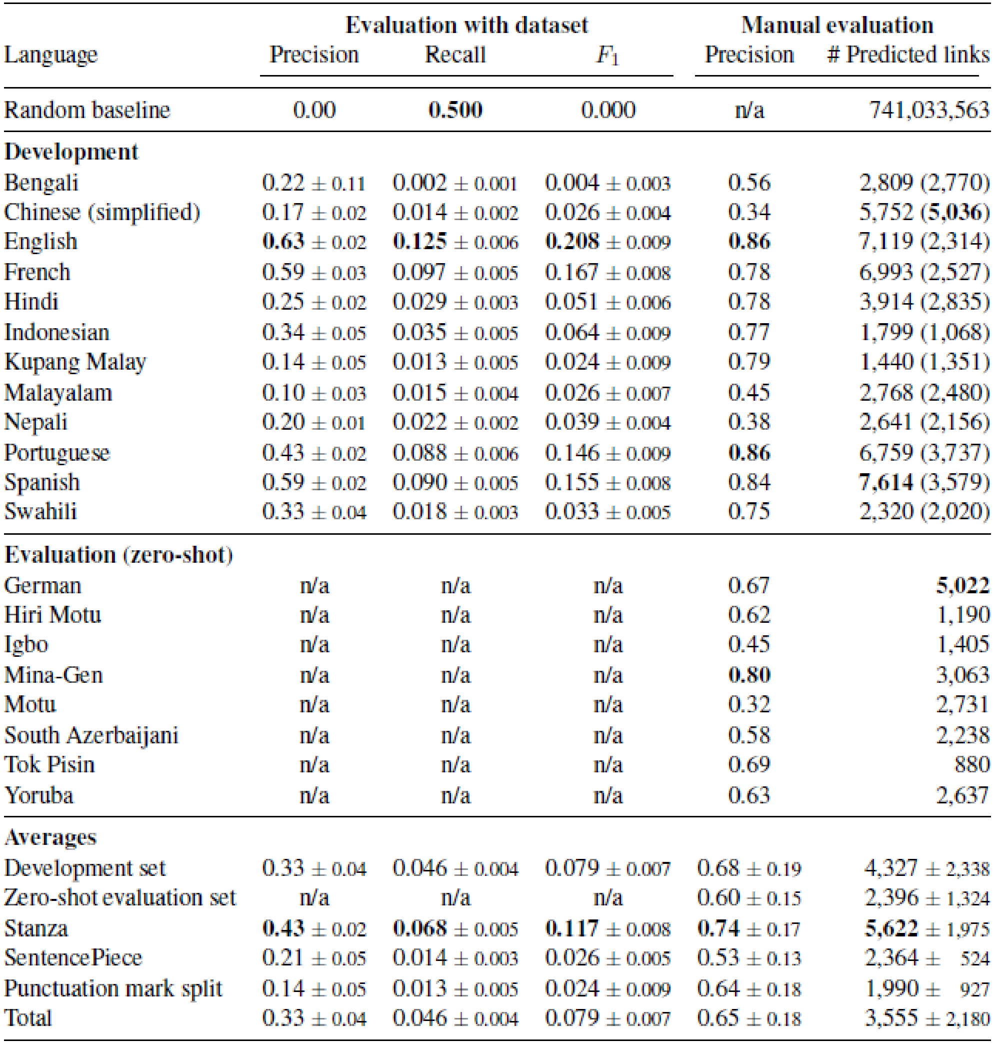


Please also answer "yes" if there is a typo but you still recognize a matching word.

context	question	word	answer
Military organization	What types of military units are there?	detachment	yes
Wrong, unsuitable	What words refer to something being unsuitable for a particular place?	discordant	yes
Hair	What words describe types of hair?	thin	yes
Sexual relations	What general words refer to sexual relations?	sex	yes
Strong	What words describe a person who is strong?	manly	yes
Work hard	What words describe someone who works too hard?	overdrive	no
Right, left	What words refer to the left side?	left	yes
Occupation	What are the occupations in manufacturing?	blacksmith	yes

- We evaluated GUIDE’s precision with **20 questionnaires**.

Results



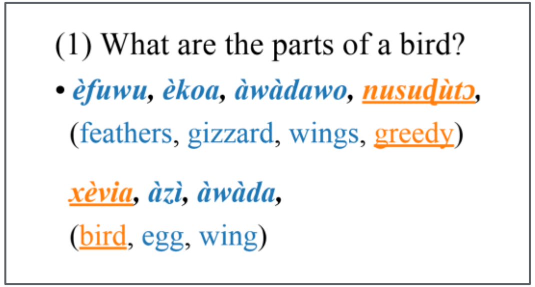
Language	Evaluation with dataset			Manual evaluation	
	Precision	Recall	F_1	Precision	# Predicted links
Random baseline	0.00	**0.500**	0.000	n/a	741,033,563
Development					
Bengali	0.22 ± 0.11	0.002 ± 0.001	0.004 ± 0.003	0.56	2,809 (2,770)
Chinese (simplified)	0.17 ± 0.02	0.014 ± 0.002	0.026 ± 0.004	0.34	5,752 (**5,036**)
English	**0.63** ± 0.02	**0.125** ± 0.006	**0.208** ± 0.009	**0.86**	7,119 (2,314)
French	0.59 ± 0.03	0.097 ± 0.005	0.167 ± 0.008	0.78	6,993 (2,527)
Hindi	0.25 ± 0.02	0.029 ± 0.003	0.051 ± 0.006	0.78	3,914 (2,835)
Indonesian	0.34 ± 0.05	0.035 ± 0.005	0.064 ± 0.009	0.77	1,799 (1,068)
Kupang Malay	0.14 ± 0.05	0.013 ± 0.005	0.024 ± 0.009	0.79	1,440 (1,351)
Malayalam	0.10 ± 0.03	0.015 ± 0.004	0.026 ± 0.007	0.45	2,768 (2,480)
Nepali	0.20 ± 0.01	0.022 ± 0.002	0.039 ± 0.004	0.38	2,641 (2,156)
Portuguese	0.43 ± 0.02	0.088 ± 0.006	0.146 ± 0.009	**0.86**	6,759 (3,737)
Spanish	0.59 ± 0.02	0.090 ± 0.005	0.155 ± 0.008	0.84	**7,614** (3,579)
Swahili	0.33 ± 0.04	0.018 ± 0.003	0.033 ± 0.005	0.75	2,320 (2,020)
Evaluation (zero-shot)					
German	n/a	n/a	n/a	0.67	**5,022**
Hiri Motu	n/a	n/a	n/a	0.62	1,190
Igbo	n/a	n/a	n/a	0.45	1,405
Mina-Gen	n/a	n/a	n/a	**0.80**	3,063
Motu	n/a	n/a	n/a	0.32	2,731
South Azerbaijani	n/a	n/a	n/a	0.58	2,238
Tok Pisin	n/a	n/a	n/a	0.69	880
Yoruba	n/a	n/a	n/a	0.63	2,637
Averages					
Development set	0.33 ± 0.04	0.046 ± 0.004	0.079 ± 0.007	0.68 ± 0.19	4,327 ± 2,338
Zero-shot evaluation set	n/a	n/a	n/a	0.60 ± 0.15	2,396 ± 1,324
Stanza	**0.43** ± 0.02	**0.068** ± 0.005	**0.117** ± 0.008	**0.74** ± 0.17	**5,622** ± 1,975
SentencePiece	0.21 ± 0.05	0.014 ± 0.003	0.026 ± 0.005	0.53 ± 0.13	2,364 ± 524
Punctuation mark split	0.14 ± 0.05	0.013 ± 0.005	0.024 ± 0.009	0.64 ± 0.18	1,990 ± 927
Total	0.33 ± 0.04	0.046 ± 0.004	0.079 ± 0.007	0.65 ± 0.18	3,555 ± 2,180

1. GUIDE has a precision of **0.65** and a recall of **0.046.**
2. The questionnaire-based precision is **twice as high** as the dataset-based precision.
3. For the zero-shot evaluation languages, GUIDE predicts **2,400** [2,020] **word-SDQ links** on average (~ 22% [21%] of the input vocabulary).
4. ⇒ GUIDE predicts **12 correct dictionary entries** for the low-resource languages in the zero-shot evaluation set per 100 words in the vocabulary.

Conclusion

Conclusion

- GUIDE = tool to **create dictionaries** in low-resource languages



(1) What are the parts of a bird?

- ***èfuwu***, ***èkoa***, ***àwàdawo***, ***nusuɖùtɔ***,
  (feathers, gizzard, wings, greedy)

  xèvia, ***àzì***, ***àwàda***,
  (bird, egg, wing)

- eBible corpus + SIL’s semantic domain dictionaries + Eflomal
  = Labeled MAG

- Labeled MAG + GCN = **33,000 correct and new** dictionary entries in 20 languages

- **Limitations:** incorrect predictions, missing predictions

- GUIDE is a **copilot** for language experts.

References

[Åkerman23] V. Åkerman, D. Baines, D. Daspit, U. Hermjakob, T. Jang, M. Martin, J. Mathew and M. Schwarting. *The eBible Corpus: Data and Model Benchmarks for Bible Translation for Low-Resource Languages.* In: arXiv (2023)

[Eberhard23] D. M. Eberhard, G. F. Simons and C. D. Fennig. *Ethnologue: Languages of the World*. Twenty-sixth edition. SIL International, 2023

[Haddow22] B. Haddow, R. Bawden, A. V. M. Barone, J. Helcl and A. Birch. *Survey of Low-Resource Machine Translation*. In: COLING 48.3 (1 Sept. 2022)

[Whistler23] K. Whistler. A Unicode Standard Annex (UAX) #15. Technical Report 15. 2023

[Imani21] A. Imani Googhari, M. Jalili Sabet, L. K. Senel, P. Dufter, F. Yvon and H. Schütze. *Graph Algorithms for Multiparallel Word Alignment*. In: EMNLP. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, 2021

[Moe10] R. Moe. *Compiling Dictionaries Using Semantic Domains*. In: Lexikos (2010)

[Scanell07] K. P. Scannell. The Crúbadán Project: *Corpus building for under-resourced languages*. In: SIGWAC. 2007

[Wu21] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang and P. S. Yu. *A Comprehensive Survey on Graph Neural Networks.* In: IEEE Transactions on Neural Networks and Learning Systems