Language Technology Lab

Towards Equitable and Culturally Adapted Multilingual Dialog Systems

Ivan Vulić
LTL, University of Cambridge

UNIVERSITY OF CAMBRIDGE

UNLP Workshop
(Online)
May 25 2024

Why Multilingual NLP?

Speaking **more languages** means communicating with **more people**…
…and reaching **more users and customers**…

Why Multilingual NLP and Multilingual Dialogue?

…but there are **more profound** and **democratic** reasons to work in this area:

- decreasing **the digital divide**
- dealing with **inequality of information (access)**
- mitigating **cross-cultural biases**
- deploying language technology for **underrepresented** languages, dialects, minorities; societal impact
- understanding cross-linguistic differences

“95% of all languages in use today will never gain traction online” (Andras Kornai)

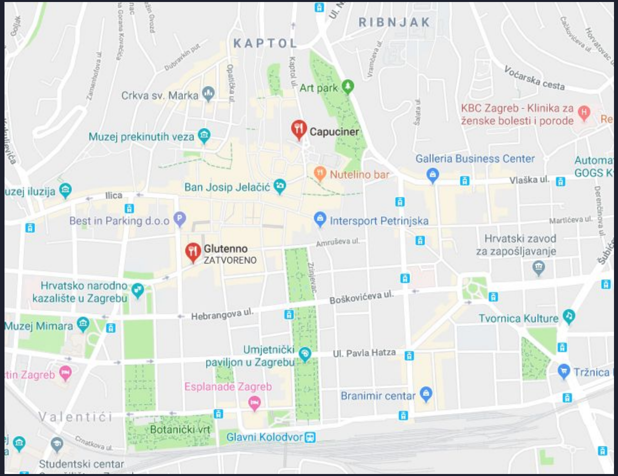
“The limits of my language *online* mean the limits of my world?”

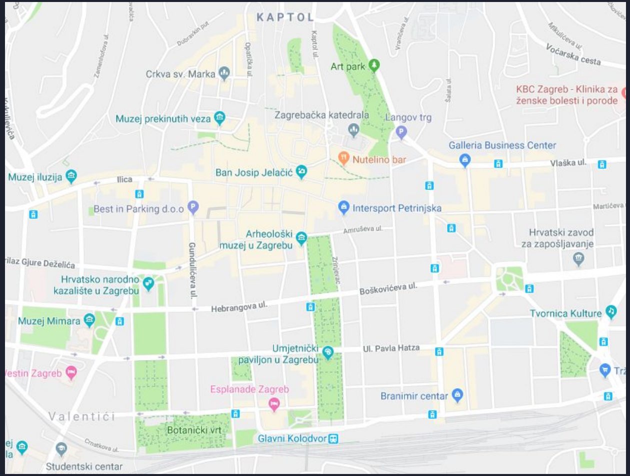
Source: http://labs.theguardian.com/digital-language-divide/

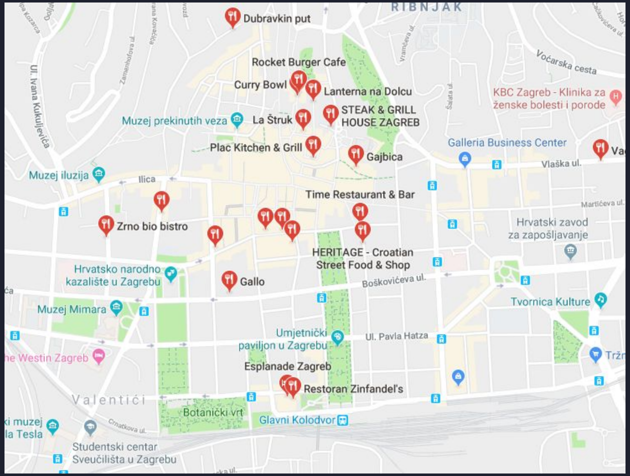
Why Multilingual NLP?

Inequality of information and representation can also affect how we understand places, events, processes...

We're in Zagreb searching for...





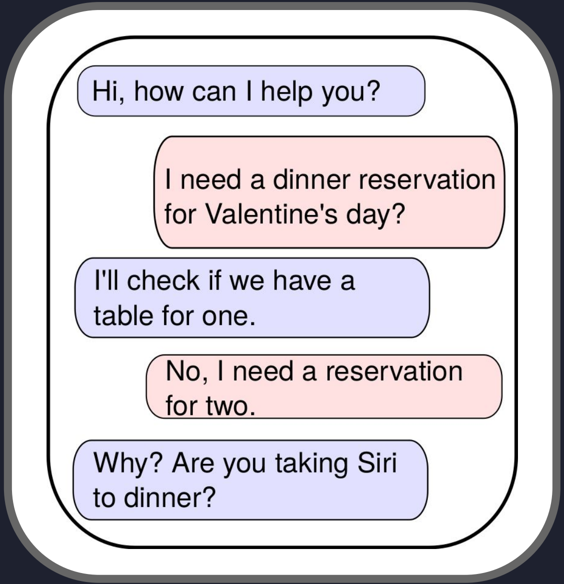


...éttermek (HU) **...jatetxea (EU)** **...restaurants (EN)**

English Dialogue Systems

A successful conversational agent must (at least implicitly) perform:

- **Automatic speech recognition (ASR)**
- **Language analysis:**
  - Language modeling, spelling correction
  - Syntactic analysis: POS tagging, parsing
  - Semantic analysis: named entity recognition, event detection, semantic role labeling, WSD
  - Coreference resolution, entity linking, commonsense reasoning, world knowledge
- **Dialog modeling:**
  - Natural language understanding, intent detection, language generation, dialog state tracking
- **Information Search and QA**
- **Text-to-Speech**



Multilingual Dialogue Systems?

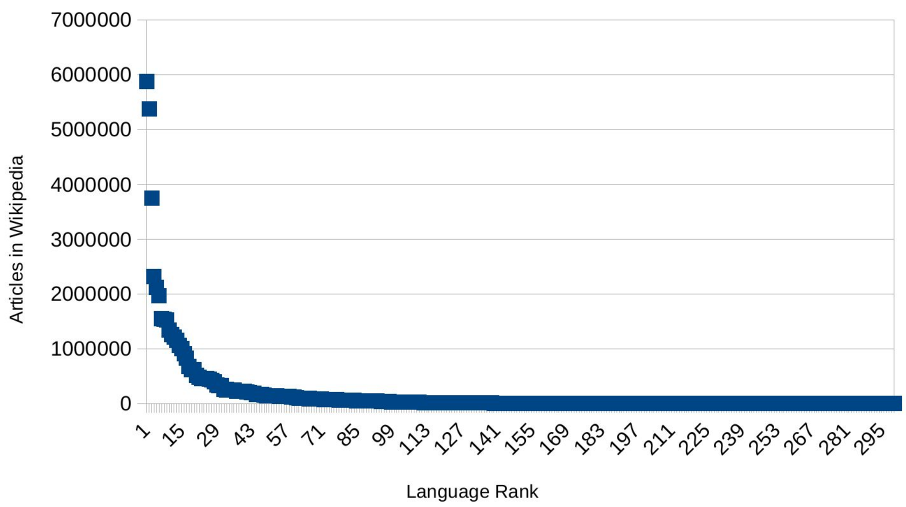
According to Ethnologue there are **7,000+** living languages

What about **language varieties** and **dialects**?

What about **“social media” languages** and **slang**?

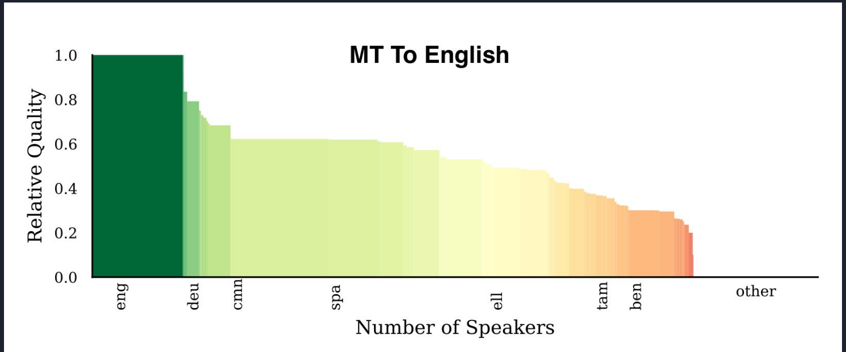
What about **“all those domains”**?

The Long Tail of Data

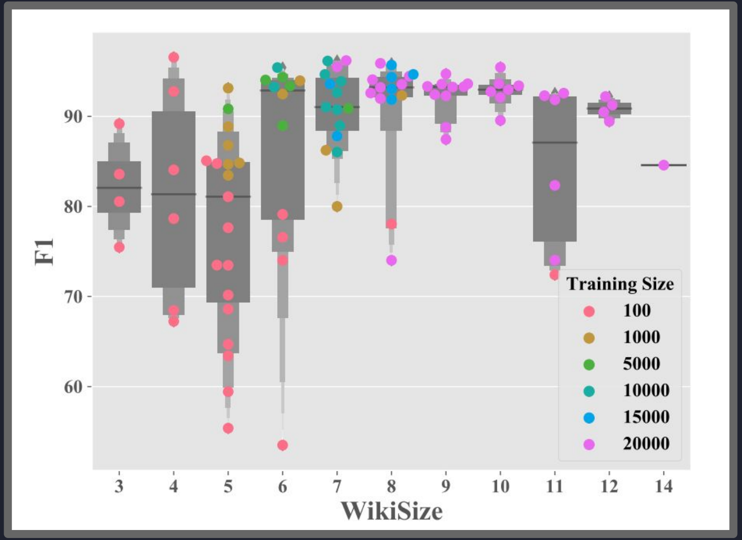


Even getting “raw” unannotated data is problematic for many languages…

The Long Tail of Data Means Inequality

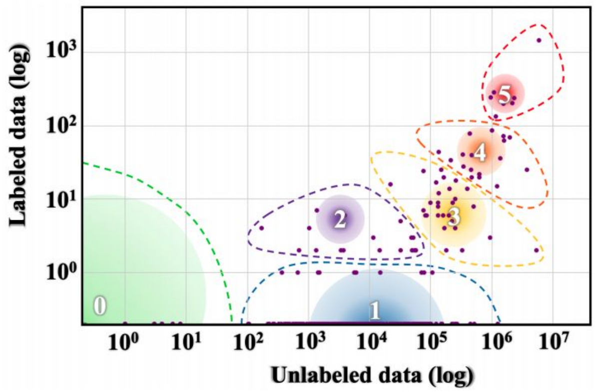


MT for major versus minor languages
(Blasi et al, 2022)



NER with mBERT on 99 languages
(Wu and Dredze, 2020)

Are All Languages Created Equal?



Most languages are "Left-Behinds"
[Joshi et al., ACL-20; Blasi et al., ACL-22]

Is creating equitable language technology across different languages then even possible?

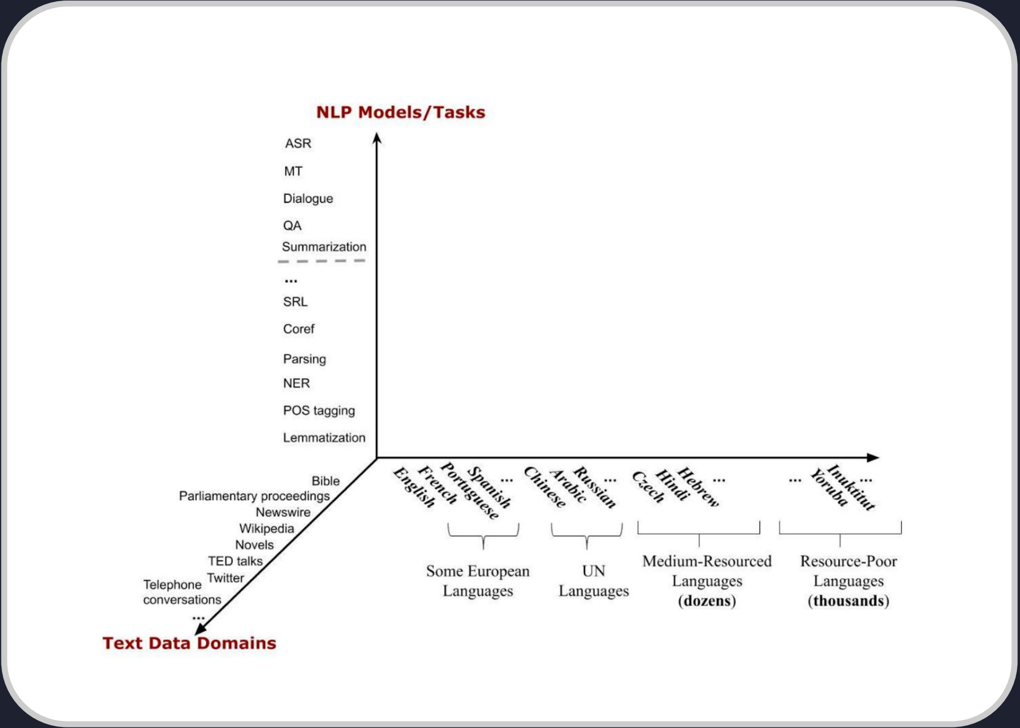
Can we at least try to 'approximate' equality?

Class	5 Example Languages	#Langs	#Speakers	% of Total Langs
0	Dahalo, Warlpiri, Popoloca, Wallisian, Bora	2191	1.2B	88.38%
1	Cherokee, Fijian, Greenlandic, Bhojpuri, Navajo	222	30M	5.49%
2	Zulu, Konkani, Lao, Maltese, Irish	19	5.7M	0.36%
3	Indonesian, Ukranian, Cebuano, Afrikaans, Hebrew	28	1.8B	4.42%
4	Russian, Hungarian, Vietnamese, Dutch, Korean	18	2.2B	1.07%
5	English, Spanish, German, Japanese, French	7	2.5B	0.28%

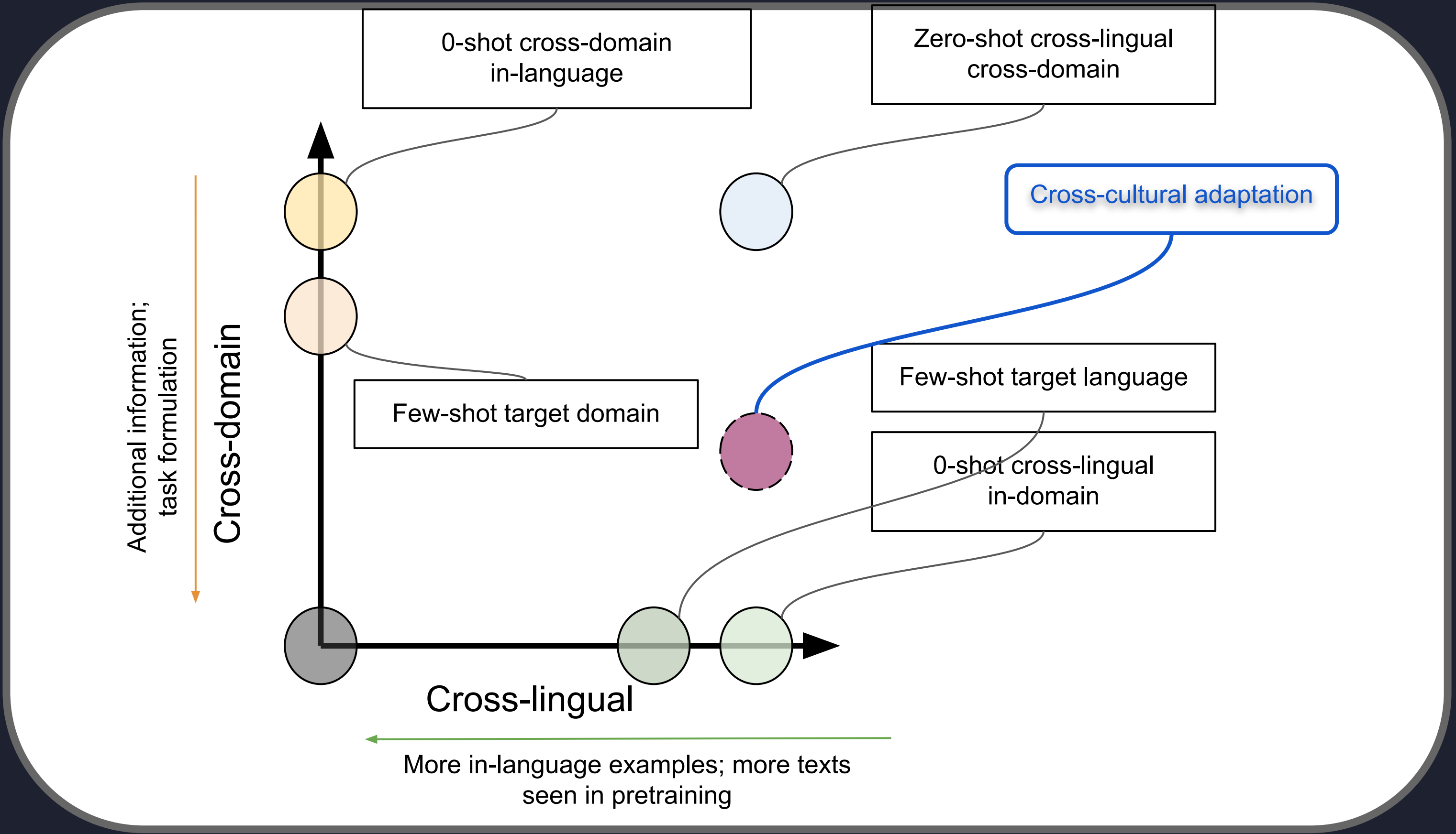
What Can We Do in Multilingual and Multi-Domain Setups?

Image courtesy of *Yulia Tsvetkov*

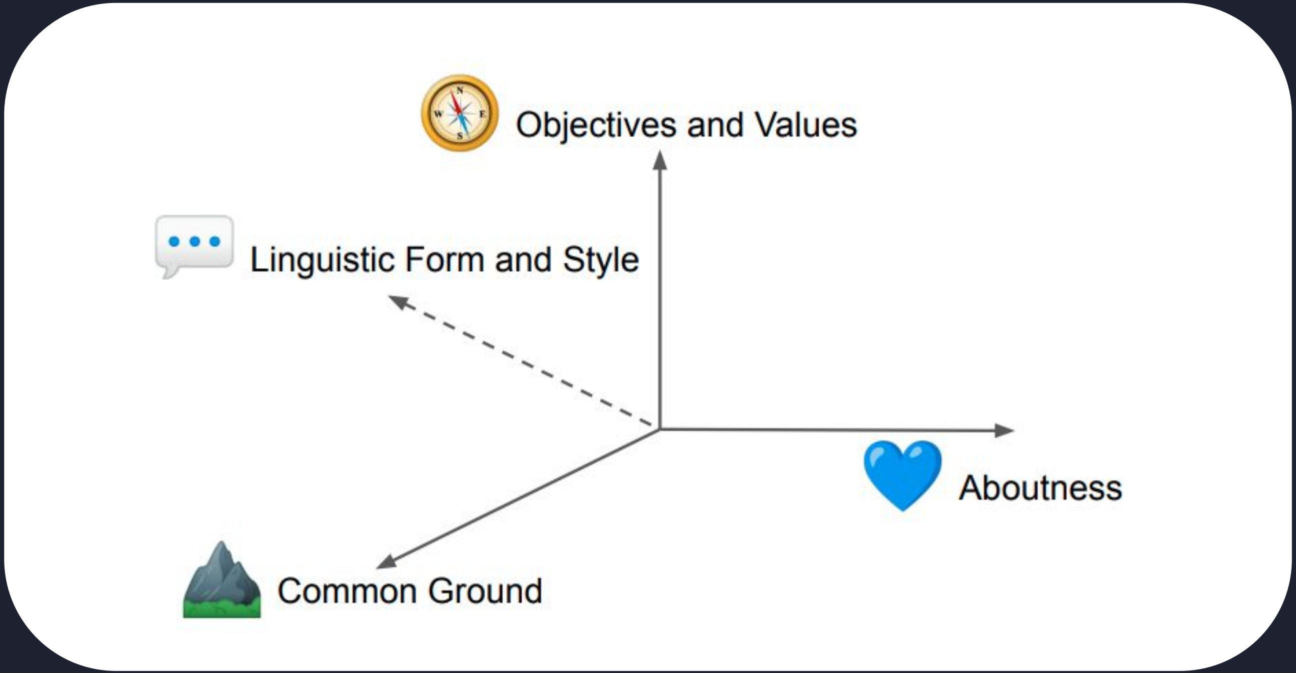
- Many NLP tasks and domains share common knowledge about language (e.g. linguistic representations, structural similarities)
- Languages and domains share common structure (on the lexical, syntactic, and semantic level)
- Annotated data is rare, make use of as much supervision as available
- Empirically, transfer learning has resulted in SOTA for many supervised NLP tasks (e.g. classification, information extraction, QA, etc)



Towards Multilingual and Multi-Domain Systems?



Why Cultural Adaptation?



[Hershcovich+, ACL-22]

Common Ground: shared / common sense knowledge
Aboutness: what people care to convey and talk about

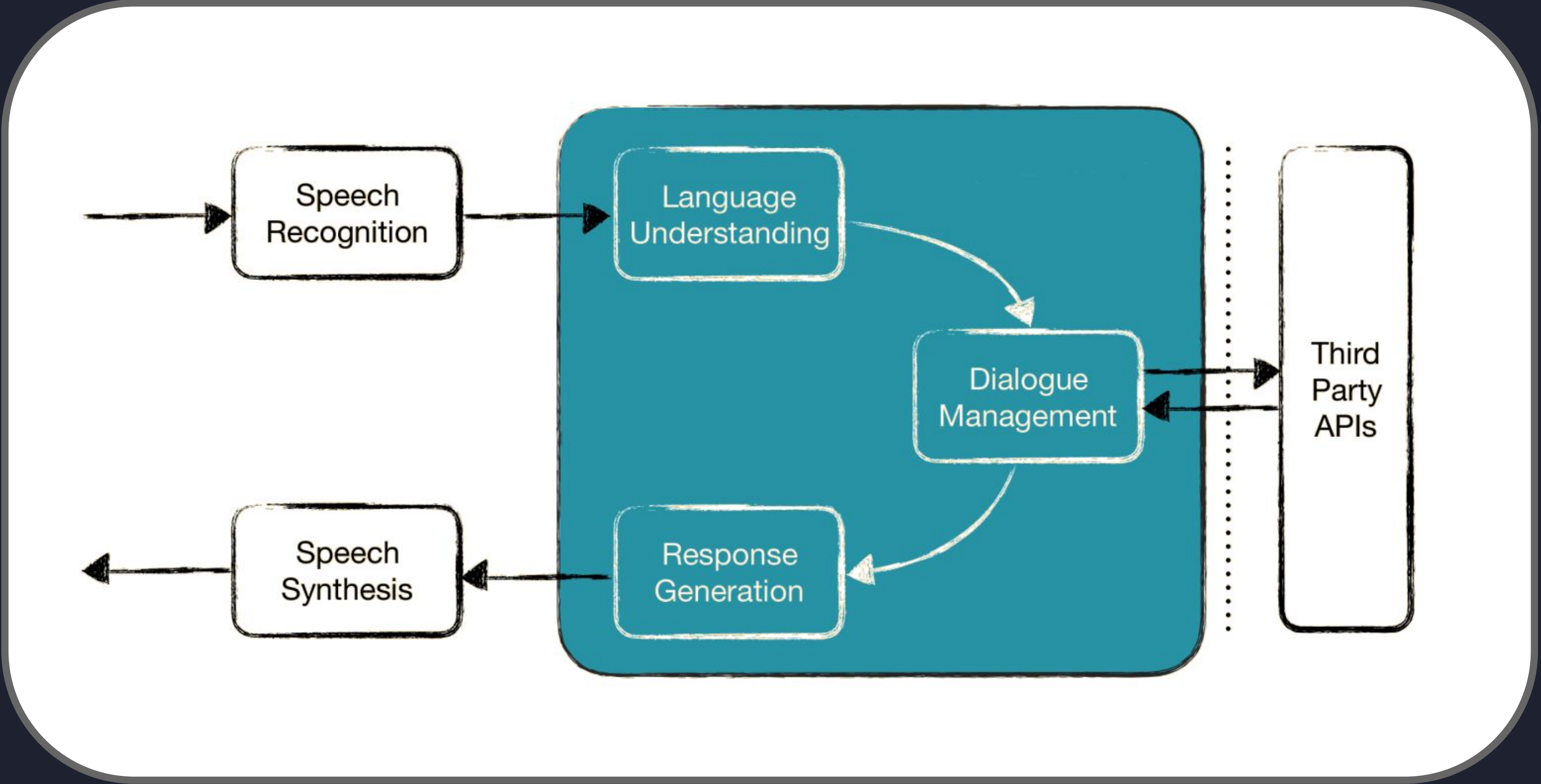
Why Cultural Adaptation?

- Mitigating conversational **bias towards source-culture concepts and contexts**
  - *Tailgating in Germany?*
  - *Baseball discussions in Croatia?*
  - *“March Madness” in Turkey?*

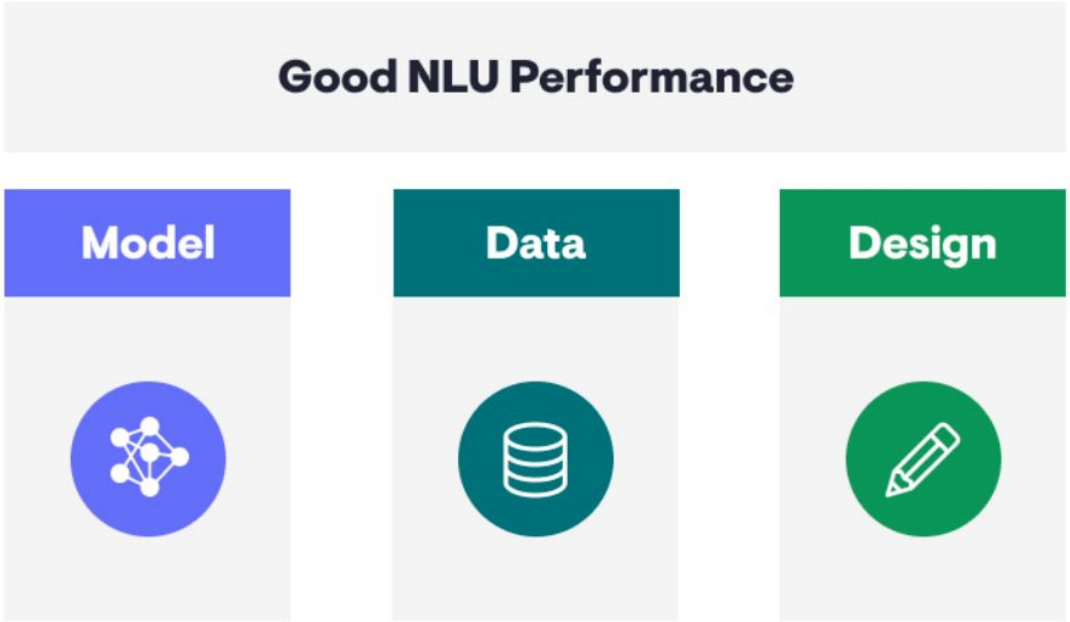
- Taking into account **specificities of the target culture**
  - *Postcode patterns vs no postcodes at all?*
  - *Buses vs trains in public transport?*

- **Avoiding `atypical’ culturally ungrounded dialogues?**
  - *The concept of “gastropub” in Arabic-speaking countries?*

("Old School") Task-Oriented Dialogue Systems



Preliminaries: Three Pillars of Dialogue (or any ML-Driven Tech)



Model

+ More efficient data usage
+ Less annotation effort
+ More accurate predictions

- Research-intensive and experimentation-intensive

Data

+ Always improves performance

- Expensive to create
- Requires multiple cycles

Design

+ Improves performance
+ Improves efficiency and 'semantic support'

- Requires some expertise, intuition and practice

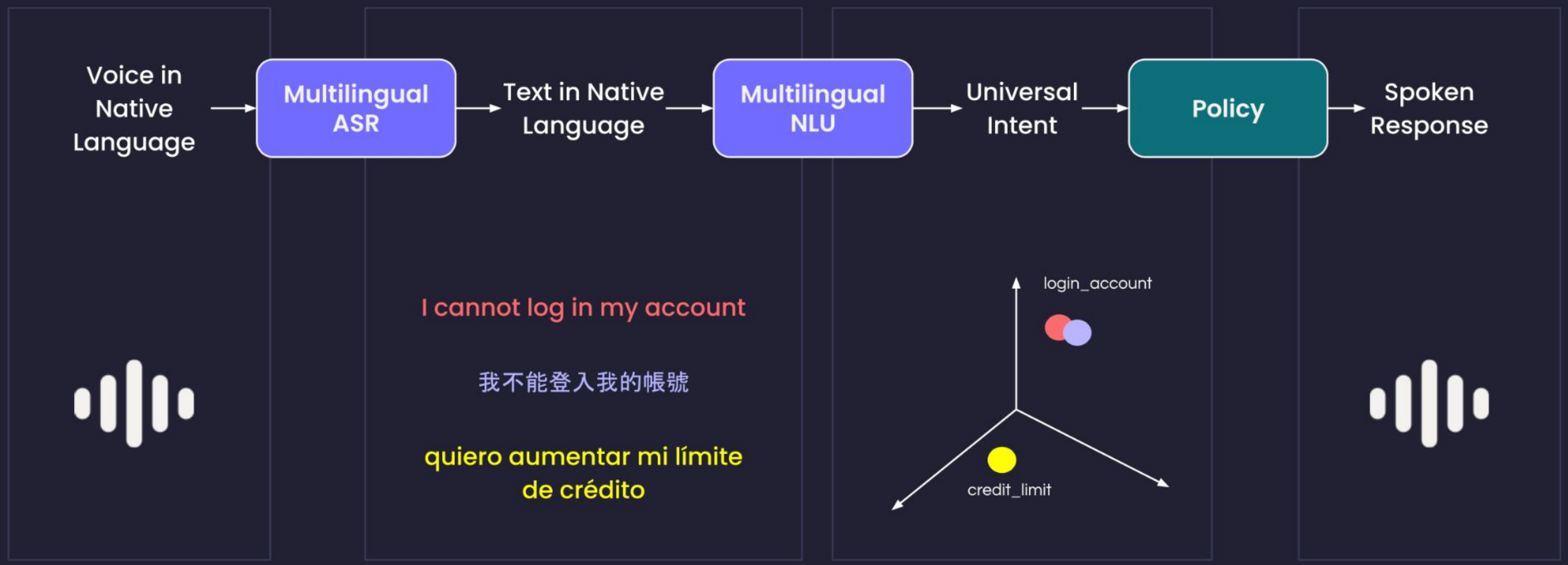
Slide courtesy of: Inigo Casanueva

MULTI^{3}NLU^{++}: A Multilingual, Multi-Intent, Multi-Domain Dataset for Natural Language Understanding in Task-Oriented Dialogue

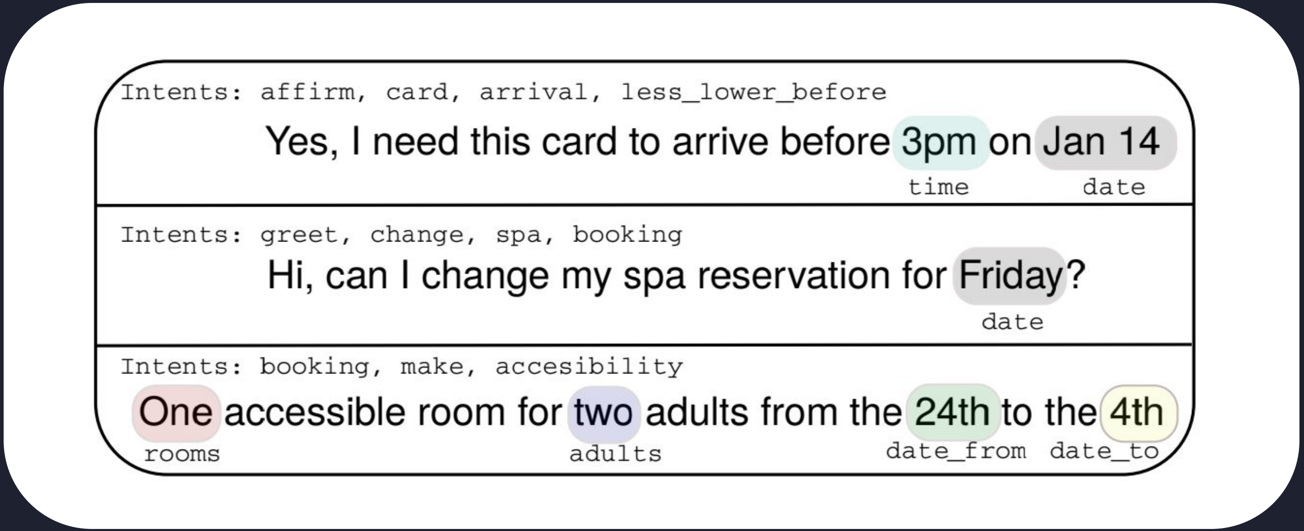
Nikita Moghe*[1]*, Evgeniia Razumovskaia*[2], Liane Guillou[1], Ivan Vulić[2], Anna Korhonen[2], Alexandra Birch[1]
School of Informatics, University of Edinburgh[1]
Language Technology Lab, University of Cambridge[2]

[Findings of ACL-23]

(Multilingual) Intent Detection



Multi-Label Intent Detection and Modular "Subintents"



- Reusability and composability **(across domains)**
- **"Semantic sharing"** and data-efficient generalisation
- Handling more complex scenarios (and with smaller intent sets)

(English-only) NLU++ [Casanueva et al., NAACL-HLT 2022]

Why Multi^{3}NLU++?

Challenges:

- Enabling training and evaluation of multi-domain NLU models in multiple languages

- When data is scarce, it should be high quality

- Cross-lingual approaches should ideally make improvements **across the 'resourceness' spectrum**

- No datasets for effective **cross-domain** and **cross-lingual** evaluation in **realistic dialogue problems**

What is Multi^{3}NLU++?

Multi^{3}NLU++, a dataset for dialogue NLU which is:

- **Multi**-lingual: English, Spanish, Turkish, Marathi, Amharic
- **Multi**-domain: BANKING and HOTELS (and COMBINED)
- **Multi-**intent: each example is labelled with multiple intents

*(and also **Multi**-parallel, it should have been called **Multi^{4}NLU++**)*

- Realistic, conversational language

As a benchmark, Multi3NLU++ allows for systematic, controlled comparison:

- across the languages with different levels of resources
- across domains – on seen and unseen intents
- across dialogue NLU tasks

What is in Multi³NLU++?

Intents: balance, overdraft, how_much

en:	I spent $58 in overdraft. What is my current balance?
am:	ከሂሳቤ ላይ ማውጣት ከምችለው በላይ 58 ዶላር አውጥቻለሁ አሁን ያለኝ ቀሪ ሂሳብ ስንት ነው?
mr:	मी ओव्हरड्राफ्टमध्ये ५८ डॉलर्स खर्च केले. माझी सध्याची शिल्लक किती आहे?
tr:	ek hesabımdan 58 dolar harcadım. Mevcut bakiyem ne?
es:	Gasté 58 dólares en descubierto. ¿Cuál es mi saldo actual?

amount_of_money

How was Multi³NLU++ Created?

Language selection: diverse level of ‘resourceness’, different language families, scripts and geographical spread

- **Spanish** - high-resource, Romance, Latin script
- **Marathi** - mid-resource, Indo-Aryan, Devanagari script
- **Turkish** - mid-resource, Turkic, Latin script
- **Amharic** - low-resource, Semitic, Ge’ez script

(p.s. Turkish is agglutinative vs synthetic/fusional languages)

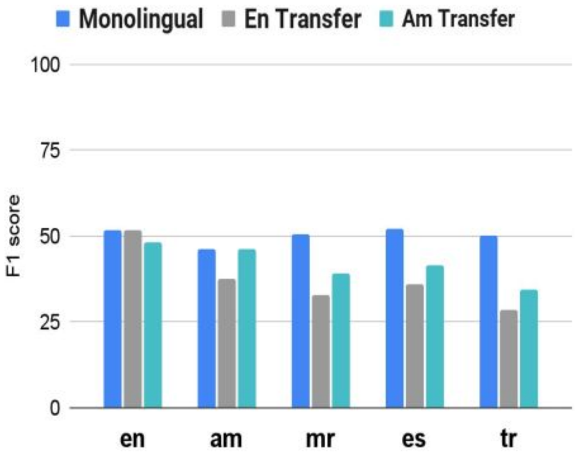
Manual translation:
Professional, aimed at preserving the colloquial nature of the English utterances

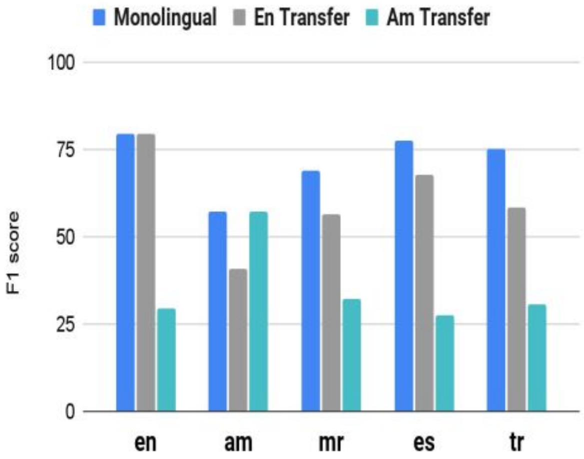
- *Three translators per language*
- *Slot span verification: IAA of ~90%*

Some Modeling Paradigms in Comparison

+ (Then) state-of-the-art sentence encoders (e.g., LaBSE)
+ (Still) state-of-the-art multilingual encoders (e.g., XLM-R, mDeBERTa)
+ We also test **standard (full) fine-tuning**

A (Tiny) Sample of Results...





- **QA-based models** are better than MLP-based models and full fine-tuning
- **Performance drop for all languages beyond English** (although the training and test data are multi-parallel!)
- **Amharic (MLP)** and **English (QA)** are best sources for cross-lingual transfer

...with Some (More) General Empirical Findings

(there are many more in the paper)

- **Overall trend across languages:** with more training data we gain better performance both in-domain and cross-domain

- **Language-specific:** absolute numbers are indicative of resources available for pretraining of the model

- **The gap between low- and high-resource languages** is rooted in (i) the amount of in-task training data; (ii) representational power of multilingual models

---

- In the **cross-domain setup** high-resource languages benefit more from the increase in training data size than lower-resource languages
- **Cross-domain cross-lingual generalisations:** The lower-resource the language, the lower the performance

Cross-Lingual Dialogue Dataset Creation via Outline-Based Generation

Olga Majewska[◇] **Evgeniia Razumovskaia**[◇] **Edoardo M. Ponti**[†◇]
Ivan Vulić[◇] **Anna Korhonen**[◇]
[◇]Language Technology Lab, University of Cambridge
[†]Institute for Language, Cognition and Computation, University of Edinburgh

[TACL-23]

Two Main Goals

1. **Get rid of translation-based and test “translation-free” data creation**

...and get rid of negative effects of “translationese”...

2. **Verify the ability to do (preliminary) cultural adaptation**

...and its importance in dialogue

(Proof-of-concept work: the total cost of the whole dataset was £800)

Bottom-Up *(Outline-Based)* Dialogue Creation

Stage 1: Source Dialogue Sampling

Stage 2: Outline Generation

Stage 3: Dialogue Writing

Stage 4: Slot Span Validation

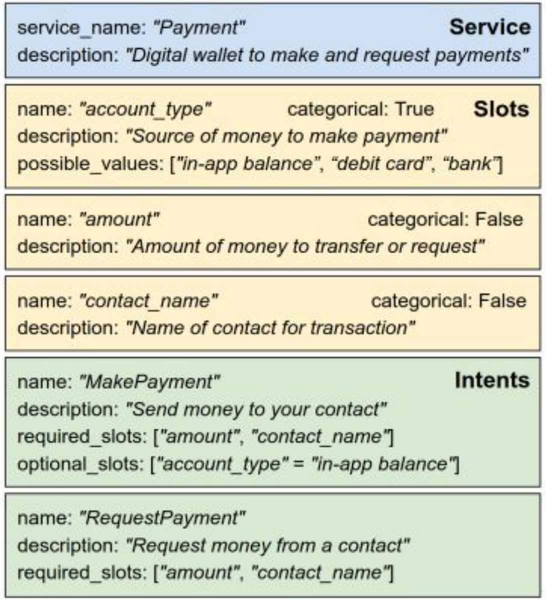
Language	ISO	Family	Branch	Macro-area	L1 [M]	Total [M]
Russian	RU	Indo-European	Balto-Slavic	Eurasia	153.7	258
Standard Arabic	AR	Afro-Asiatic	Semitic	Eurasia / Africa	$0^\dagger$	274
Indonesian	ID	Austronesian	Malayo-Polynesian	Papunesia	43.6	199
Kiswahili	SW	Niger–Congo	Bantu	Africa	16.3	69

Stage 1: Source Dialogue Sampling

Starting point: English Schema-Guided Dialogue (SGD) Dataset

- Readily available (abstracted) dialogue schemata [Rastogi et al., arXiv-19]
- We randomly sample dialogues for 11 domains, 10 examples per intent

	Alarm (◇)	Flights	Homes	Movies	Music	Media	Banks	Payment (◇)	RideSharing	Travel	Weather	#turns
Dev	13	12	12	16	14	-	14	-	-	12	18	**1138**
Test	21	23	13	19	16	17	-	8	11	-	-	**1352**



Stage 2: Outline Generation

Act	Slot/Intent	Description	Value	Outline
INFORM_INTENT	SearchOnewayFlight	Search for one-way flights to the destination of choice	–	***Express the desire to*** *search for one-way flights*
REQUEST	number_checked_bags	Number of bags to check in	2	***Ask if*** *the number of bags to check in is 2*

(Some) Cultural Adaptation happens here:

- *New York -> Jakarta*
- *American Airlines -> Garuda Indonesia*
- *$ -> Rp (Rupiah)*
- ...

	Localised Slot Values				
Split	AR	ID	RU	SW	AVG
Dev	42.98	59.68	61.60	76.51	60.19
Test	13.50	57.00	53.81	78.34	50.66

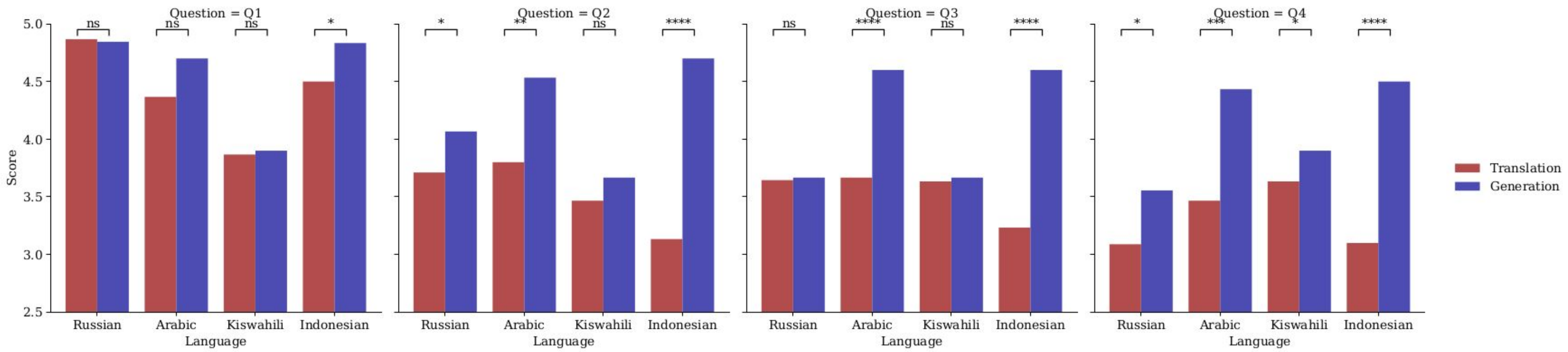
Stage 3: Dialogue Writing from Outlines

Outlines
USER: *Express the desire to search for roundtrip flights for a trip*
the name of the airport or city to arrive at: Seattle the company that provides air transport services: American Airlines
ASSISTANT/SYSTEM: *Inform the user that you found 1 such option(s). Offer the following option(s):*
the company that provides air transport services: American Airlines departure time of the flight flying to the destination: 7:35am departure time of the flight coming back from the trip: 4:15pm the total cost of the flight tickets: $343

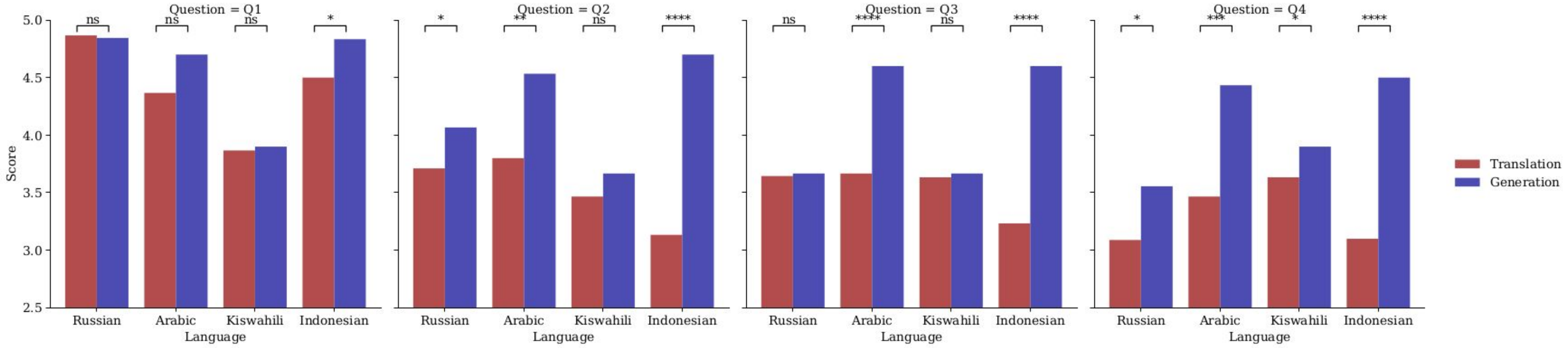
Stage 4: Slot span verification (~99% agreement)

Impact of Outline-Based Creation and Cultural Adaptation?

Questions
Q1. The ASSISTANT helps satisfy the USER's requests.
Q2. The USER speaks naturally and sounds like an Arabic native speaker.
Q3. The ASSISTANT speaks naturally and sounds like an Arabic native speaker.
Q4. I can easily imagine myself mentioning or hearing the proper names referred to in the dialogue (e.g., titles of films or songs, people, places) in a conversation with my Arabic friends or family.



Impact of Outline-Based Creation and Cultural Adaptation?

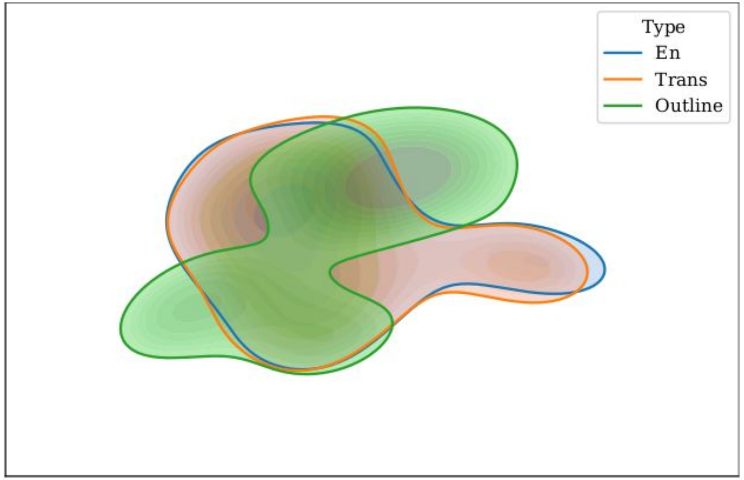


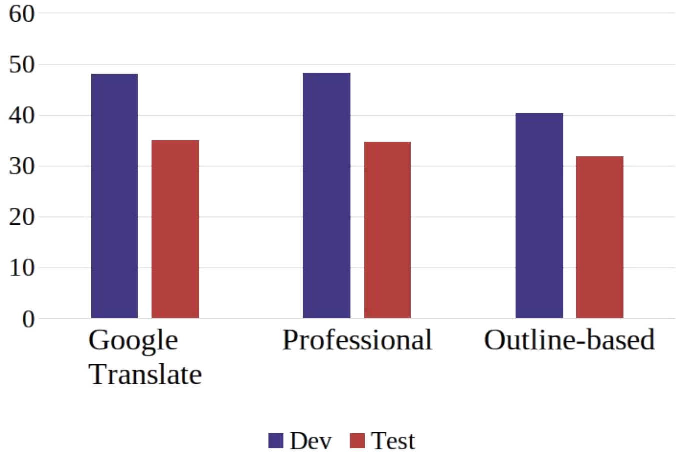
- Improved **naturalness, target-language fluency** (Q2, Q3) and **cultural familiarity** of entities (Q4)

- Effects of "translationese" in direct translation output:
  - **syntactic calques** (*"The meeting has been scheduled"*), **source lexical bias**

- **A/B Test** with 15 human participants per language: COD-based dialogues are more **natural-sounding** (80%+ in all 4 languages)

MT-Based Data Creation Inflates Performance





...and this is the consequence...

“Non-natural” alignment of data samples?

MULTI[3]WOZ: A Multilingual, Multi-Domain, Multi-Parallel Dataset for Training and Evaluating Culturally Adapted Task-Oriented Dialog Systems

Songbo Hu[1*] **Han Zhou**[1*] **Mete Hergul**[1]
Milan Gritta[2] **Guchun Zhang**[2] **Ignacio Iacobacci**[2]
Ivan Vulić[1†] **Anna Korhonen**[1†]
[1]Language Technology Lab, University of Cambridge, UK
[2]Huawei Noah's Ark Lab, London, UK

[TACL-23]

Summary of Multilingual ToD Datasets

Dataset	# Langs	# Domains	# Train	# Test	No Translation?	Culturally Adapted?	Multi-P?
WOZ 2.0	3	1	600	400	✗	✗	✓
BiToD	2	5	2,894	451	✓	✓	✗
AllWOZ	8	5	40	50	✗	✗	✓
GlobalWOZ	21	7	0 (8,437)	500 (1,000)	✗	✓	✗
Multi^{2}WOZ	5	7	0	1,000	✗	✗	✓
Multi3WOZ	4	7	7,440	860	✓	✓	✓

The need has been recognised

The solutions have been (too) quick or inadequate

Solutions Have Been (Too) Quick and Inadequate

GlobalWOZ is full of design-triggered issues and inconsistencies**:**

- **Inconsistent script-switched** and **code-switched** dialogues
- Erroneous **slot-value annotations inconsistent with dialogue ontology** and **database**
- **Contextual inconsistencies**
- **Translation-based:** automatic NMT for training data + PEMT for test data
- **Test sets** in different languages are not parallel
  - + *A heuristic for selection of dialogues penalises lexical variation*

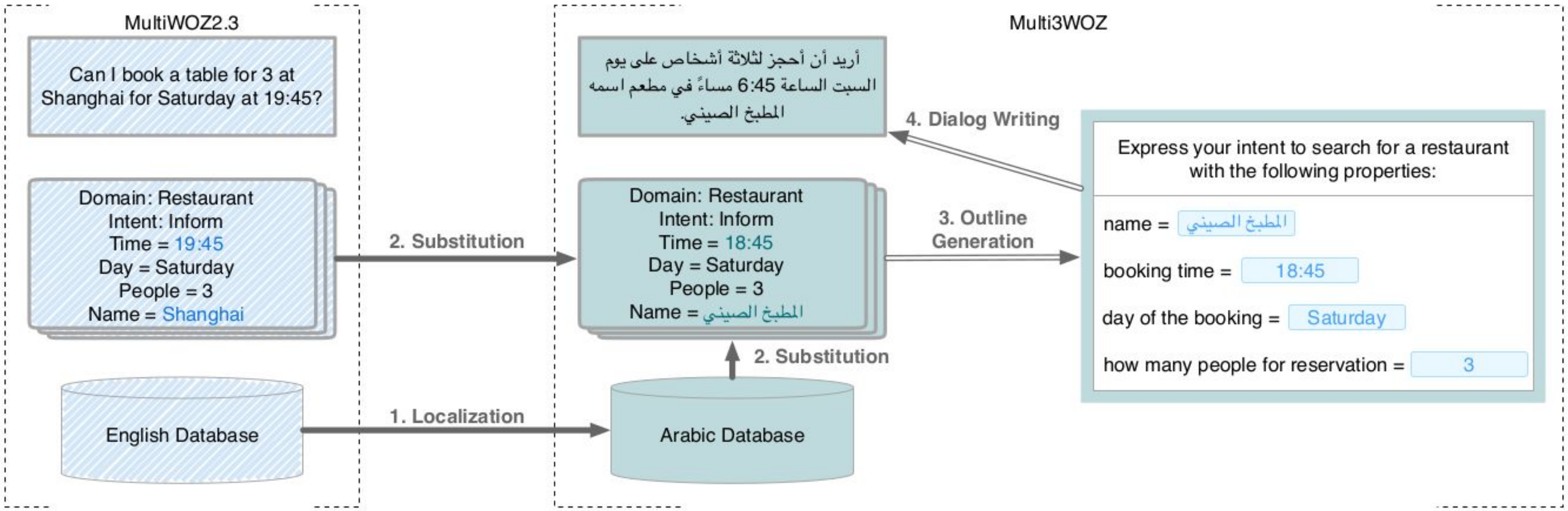
Main Goals

1. Get rid of translation-based and collect “translation-free” data multilingually **on a large scale**
- *High-quality training, dev, and test data (e.g., getting rid of GlobalWOZ-style issues)*

2. Training, evaluation, and analysis of multilingual and cross-lingual ToD systems
- ***Multi**-parallel, **multi**-lingual, **multi**-domain*

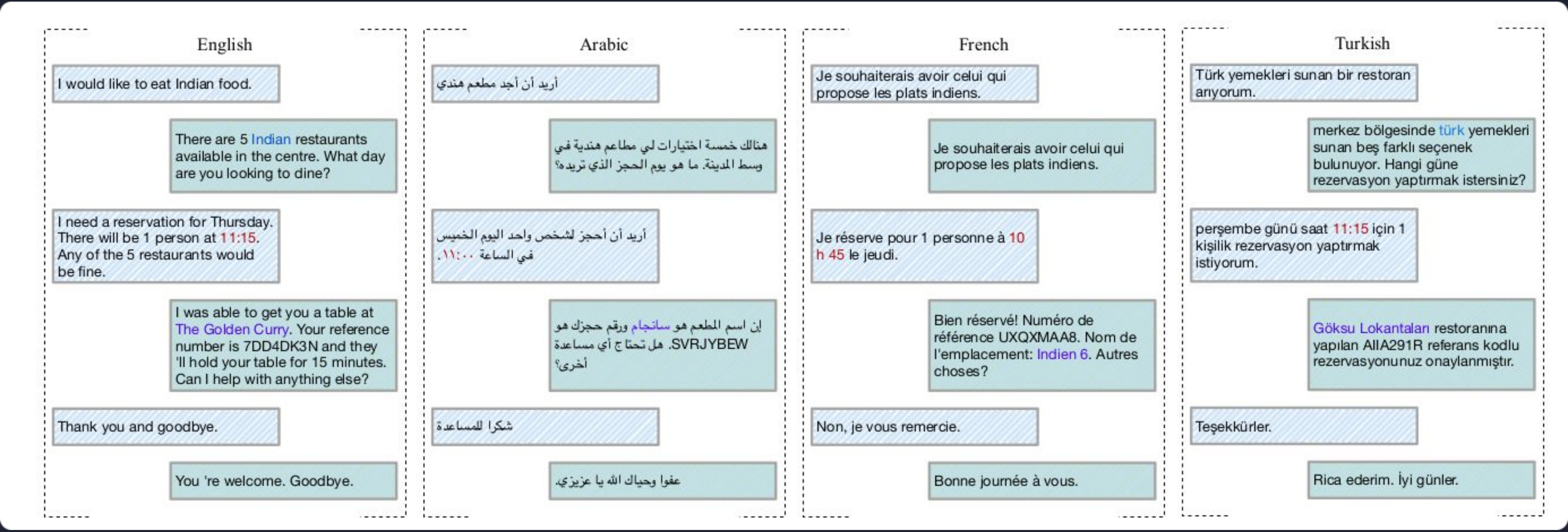
(Not a proof-of-concept any more: the total cost of the whole dataset was ~£55,000)

Multi³WOZ Dataset Construction: Bottom-Up (again)



- **"COD"-verified outline-based design!**
- **Cultural adaptation:** 1. localization, 2. substitution

Example (Parallel) Dialogues



- **Cultural adaptation:** slot-value redistribution, slot-value randomization, controlled entity replacement

A Systematic Study of Performance Disparities in Multilingual Task-Oriented Dialogue Systems

Songbo Hu[1] Han Zhou[1] Zhangdie Yuan[2] Milan Gritta[3]
Guchun Zhang[3] Ignacio Iacobacci[3] Anna Korhonen[1] Ivan Vulić[1]
[1]Language Technology Lab, University of Cambridge, UK
[2]Department of Computer Science and Technology, University of Cambridge, UK
[3]Huawei Noah's Ark Lab, London, UK

[EMNLP-23]

Multi^{3}WOZ is multi-parallel and contains abundant (high-quality) training data:

Analyses across different:

- Source and target languages (cross-lingual transfer)
- Domains (cross-domain transfer)
- Learning setups ("many"-shot vs few-shot vs zero-shot)

Preliminaries and Notation

- $P(\cdot)$: a dialogue model
- D: a task-specific dialogue dataset
- D^{src}: a typically high-resource source language dataset
- D^{tgt}: a low-resource target language dataset with equal size and quality as D^{src}
- D^{tgt}_{few}: a realistic low-resource target language dataset, which is considerably smaller compared to D^{src} and D^{tgt}
- $\mathrm{M}(\cdot)$: an automatic evaluation metric

Notions of Equivalence in Performance

- **Absolute θ-Equivalence**: we define that two systems achieve absolute θ-equivalence iff $\mathrm{M}(P^{tgt}(\cdot)) \geq \theta \cdot \mathrm{M}(P^{src}(\cdot))$, where $\theta \in [0, 1]$.
- **Relative θ-Equivalence**: We define that the two systems achieve relative θ-equivalence iff the metric $\mathrm{M}(P^{tgt}_{few}(\cdot)) \geq \theta \cdot \mathrm{M}(P^{tgt}(\cdot))$, where $\theta \in [0, 1]$.

(RQ1) Supervised, Translation-Based, Zero-Shot

RQ1) Given recent progress in multilingual LMs, machine translation, and cross-lingual transfer, is language-specific data still necessary for the development of a T O D system for a new language?

(RQ1) Supervised, Translation-Based, Zero-Shot

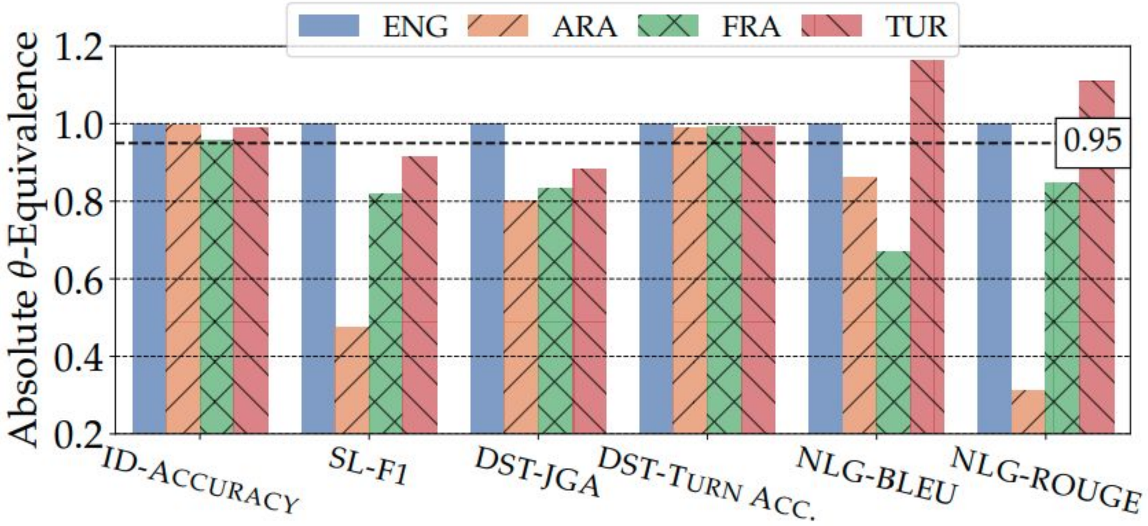
Language	Intent Detection		Slot Labelling			Dialogue State Tracking			Natural Language Generation		
	Accuracy	F1	Precision	Recall	F1	JGA	Turn Acc.	F1	BLEU	ROUGE	METEOR
Fully Supervised											
ENG	$93.2_{92.0}$	$96.1_{95.3}$	$94.6_{93.6}$	$95.7_{96.0}$	$95.1_{94.8}$	$57.2_{59.8}$	$97.7_{97.9}$	$92.5_{93.5}$	$20.1_{20.8}$	$47.3_{48.4}$	$42.9_{44.1}$
ARA	$92.7_{92.1}$	$95.0_{94.6}$	$42.4_{42.2}$	$48.5_{48.1}$	$45.2_{45.0}$	$42.0_{47.9}$	$96.4_{96.9}$	$88.0_{89.4}$	$6.8_{17.9}$	$0.8_{15.0}$	$19.4_{36.0}$
FRA	$89.2_{88.6}$	$93.0_{92.6}$	$76.9_{77.1}$	$79.2_{79.1}$	$78.0_{78.1}$	$47.6_{49.7}$	$96.8_{97.0}$	$89.4_{90.1}$	$12.9_{13.9}$	$39.6_{40.9}$	$33.8_{35.2}$
TUR	$92.2_{91.5}$	$95.0_{94.4}$	$76.9_{77.1}$	$87.6_{87.3}$	$87.1_{86.9}$	$50.5_{52.9}$	$97.1_{97.3}$	$90.5_{91.2}$	$5.5_{24.2}$	$24.7_{53.7}$	$22.5_{48.6}$
Zero-shot Cross-lingual Transfer											
ARA	$82.1_{65.7}$	$88.2_{74.8}$	$27.4_{17.2}$	$31.2_{27.7}$	$29.2_{21.2}$	$1.9_{1.5}$	$82.5_{80.7}$	$17.0_{5.8}$	$0.2_{0.2}$	$2.5_{2.1}$	$2.4_{2.0}$
FRA	$83.9_{77.0}$	$89.8_{85.0}$	$58.5_{49.1}$	$61.2_{62.4}$	$59.8_{54.9}$	$5.5_{3.7}$	$86.6_{85.1}$	$40.1_{32.8}$	$0.5_{0.4}$	$4.2_{4.7}$	$6.1_{5.9}$
TUR	$87.0_{74.9}$	$91.4_{81.7}$	$68.1_{48.5}$	$74.7_{66.6}$	$71.2_{56.2}$	$3.5_{1.3}$	$85.2_{82.1}$	$34.4_{15.2}$	$0.3_{0.4}$	$3.7_{4.4}$	$6.1_{5.8}$
Translate Train											
ARA	$72.0_{67.3}$	$81.9_{78.9}$	0_{0}	0_{0}	0_{0}	$9.2_{32.4}$	$89.1_{94.2}$	$52.7_{79.9}$	$1.1_{1.2}$	$6.3_{6.7}$	$7.4_{7.6}$
FRA	$66.2_{63.4}$	$77.4_{74.9}$	0_{0}	0_{0}	0_{0}	$10.4_{9.8}$	$90.6_{90.6}$	$60.0_{58.7}$	$2.6_{3.2}$	$20.4_{23.2}$	$15.1_{17.8}$
TUR	$71.2_{66.5}$	$82.2_{78.6}$	0_{0}	0_{0}	0_{0}	$10.5_{32.9}$	$90.5_{94.3}$	$60.4_{79.7}$	$1.0_{1.0}$	$16.9_{17.4}$	$12.7_{13.0}$

In-language data is crucial for performance

(RQ2) Intrinsic Bias in Multilingual Language Models

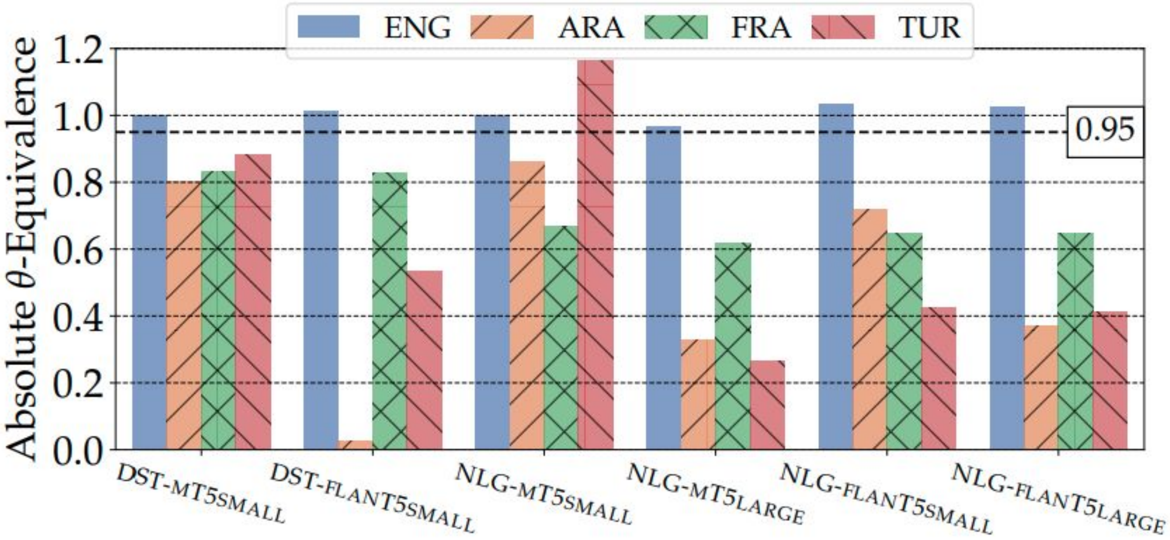
(RQ2) Given access to the same mPLMs, equivalent amounts of high-quality in-language training data, and a similar development approach as that used to create an English ToD dataset, is it possible to develop a ToD system for a new language that achieves near-English performance?

(RQ2) Intrinsic Bias in Multilingual Language Models



Intrinsic bias is prominent, but depends on task complexity as well as evaluation metrics.

(RQ2) Intrinsic Bias in Multilingual Language Models

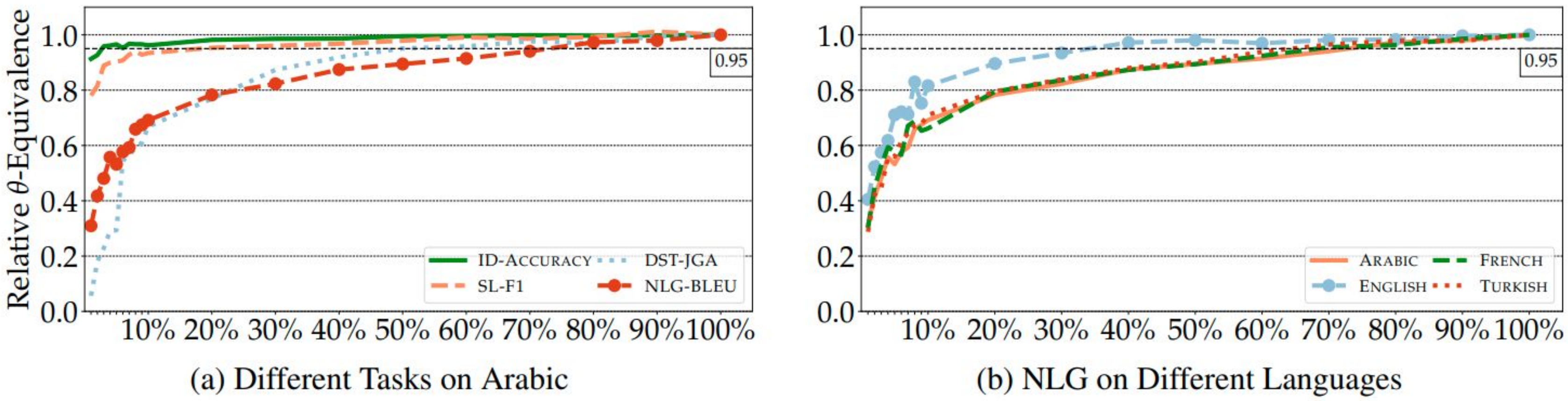


Intrinsic bias is prominent, but also depends on the chosen model, and it also exists with monolingual models

(RQ3) Adaptation Bias in Few-Shot Learning

(RQ3) How much training data is required in a new language to achieve performance comparable to a ToD system trained with an equivalent amount of in-domain, in-language data as in English?

(RQ3) Adaptation Bias in Few-Shot Learning



English is favoured even when it comes to collecting annotated task data...

(RQ4) Cost Efficiency in ToD Data Collection

(RQ4) Which data collection strategy maximises system performance across metrics while minimising the amount of annotation required? Such a strategy could optimise the cost-efficiency of annotation for a new language.

(RQ4) Cost Efficiency in ToD Data Collection

Strategy	ID	SL	DST	NLG
	Accuracy	F1	JGA	BLEU
Random Sampling	87.9	65.2	20.7	10.4
Max N-gram	**88.8**	**66.5**	23.6	**12.2**
Equal Domain	87.2	65.3	21.1	10.1
Equal Slot	87.9	65.0	26.2	11.3
Max Length	88.3	66.4	**26.7**	11.5

Averages over the three target languages based on 5% of target language data sampled/created using one of the strategies

Tip: Be less random than random sampling
Future work: Active learning? More sophisticated heuristics?

Run Your Own Experiments and Analyses

Use DiaLight [Hu et al., NAACL-24: Demos]

Toolkit	Human Evaluation	Multilinguality	LLM+E2E	Comparative Experiment
PyDial	✓	✗	✗	✗
ConvLab2	✓	✗	✗	✗
ConvLab3	✓	✓	✗	✗
to-llm-bot	✗	✗	✓	✗
other E2E baselines	✗	✗	✗	✗
DIALIGHT(this work)	✓	✓	✓	✓

DiaLight supports:

- fine-tuning and in-context learning for development
- a comprehensive and simple framework for human evaluation
- creation of interactive systems that you can chat with

UNIVERSITY OF CAMBRIDGE

Language Technology Lab

Cultural Adaptation in Dialogue is:

- *Necessary*
- *Multi-Faceted / Multi-Layered*
- *Difficult*
- *Contextual*
- *Task-Specific*
- *Underexplored*
  - *Both from data and methodology angle*

Achieving Multilingual (Performance) Equity in Dialogue is:

- *Necessary*
- *Multi-Faceted / Multi-Layered*
- *Difficult*
- *Contextual*
- *Task-Specific*
- *Underexplored*
    - *Both from data and methodology angle*

While we haven’t even scratched the surface of both. What about:

- *Low-resource languages?*
- *Non-standard language varieties?*
- *Very complex and specific domains?*
- *Proper end-to-end learning (LLMs with RAG?)*
- *Many other types of equity (and more generally DEI) beyond performance only*

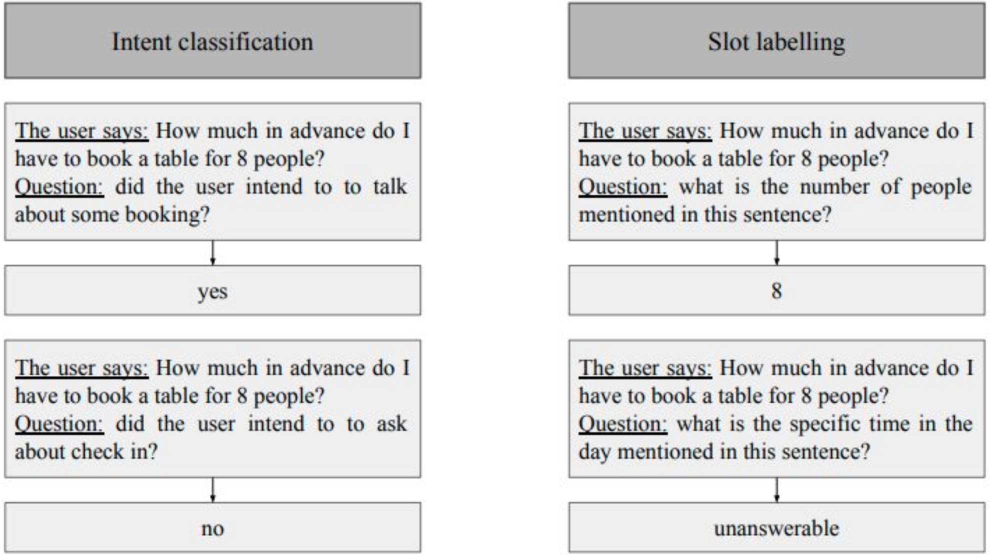
Bonus: Tackling Data Scarcity with Data-Efficient Methods

SQATIN: Supervised Instruction Tuning Meets Question Answering for Improved Dialogue NLU

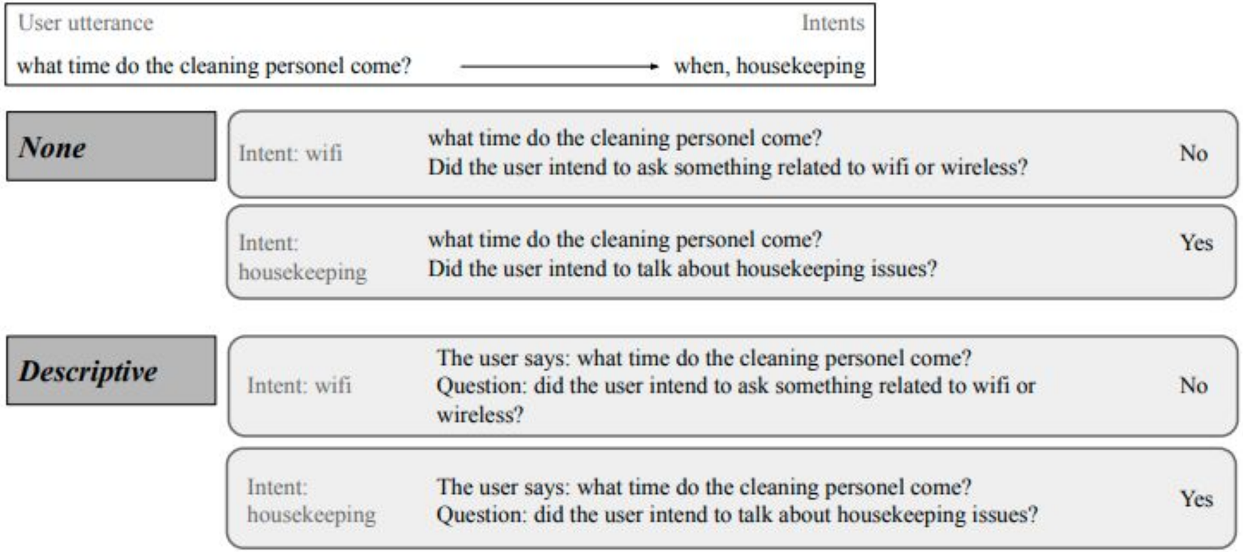
Evgeniia Razumovskaia[1], Goran Glavaš[2], Anna Korhonen[1], Ivan Vulić[1,3]
[1] Language Technology Lab, University of Cambridge
[2] Center for Artificial Intelligence and Data Science, University of Würzburg
[3] PolyAI Limited

[NAACL-24]

QA-Based Instruction Tuning of “Small LLMs”



QA-Based Instruction Tuning of “Small LLMs”



QA-Based Formulation Wins

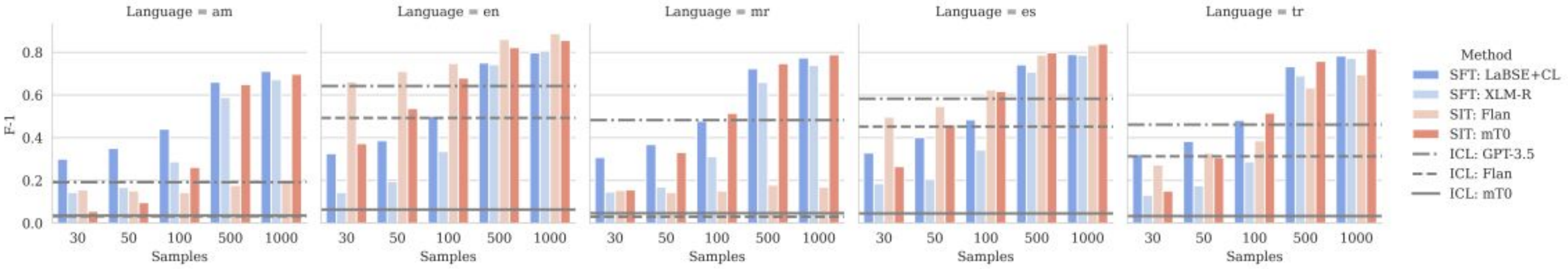
Model	Templ.	ID		VE	
		20-F	10-F	20-F	10-F
BANKING					
CL-SE		58.1	68.8	N/A	N/A
QA-FT: RoBERTa		80.3	85.6	50.5	56.7
QA-FT: mDeBERTa		80.8	85.0	59.7	66.5
QA-FT: T5		82.7	86.8	61.5	73.5
SQATIN	*None*	85.6	**88.5**	64.9	75.4
	Desc.	**85.8**	88.4	**66.3**	**76.3**
HOTELS					
CL-SE		51.9	61.8	N/A	N/A
QA-FT: RoBERTa		67.4	73.3	48.1	52.4
QA-FT: mDeBERTa		66.9	73.2	**61.6**	67.3
QA-FT: T5		69.2	76.5	57.2	**67.9**
SQATIN	*None*	73.1	78.0	58.0	**67.7**
	Desc.	**73.4**	**78.1**	58.7	67.0

The results are on English-only NLU++

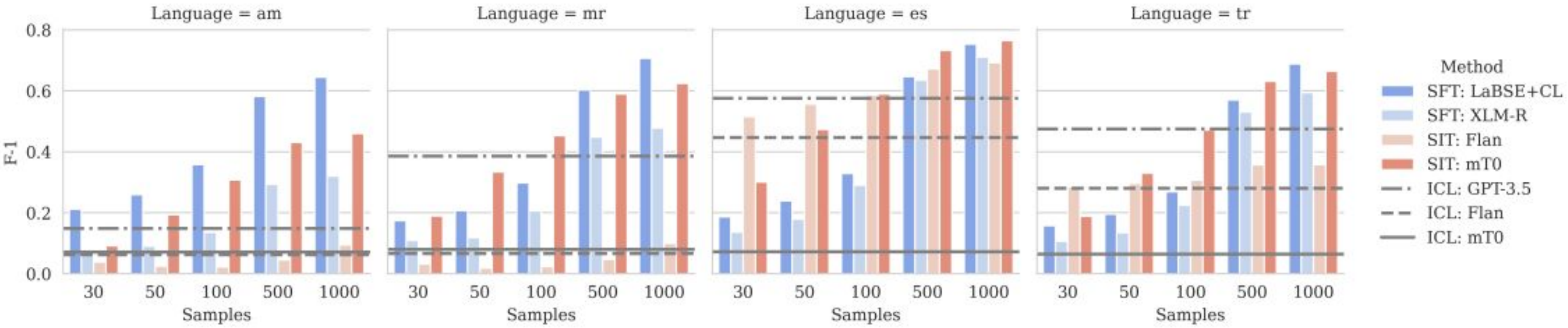
SQATIN is the most robust method in low-resource setups and across the board

Figures taken from [Razumovskaia et al., arXiv-24] *(under review in TACL)*

What about Multilingual and Cross-Lingual Setups?



(a) ID: In-Language In-Domain

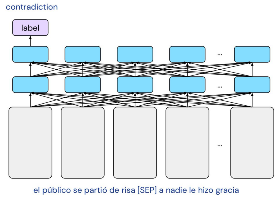


(c) ID: Cross-Lingual In-Domain

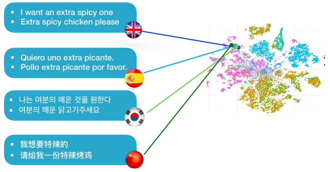
Towards Inclusive, Sustainable, Equitable Multilingual TOD

Widening the global reach of NLP: Far-reaching technological and socioeconomic consequences

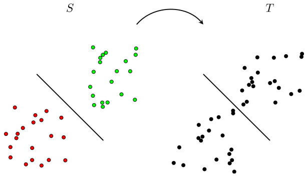
Deep (machine) learning



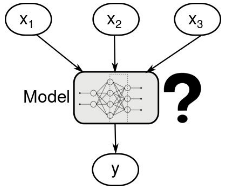
Multilingual representation learning



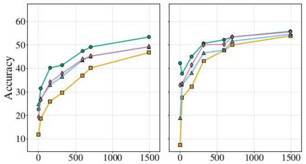
Cross-lingual knowledge transfer



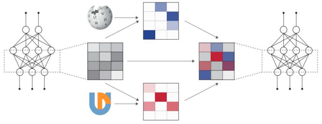
Transparency



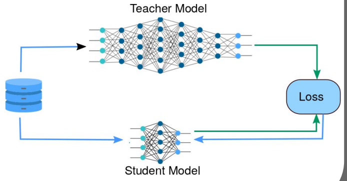
Sample Efficiency



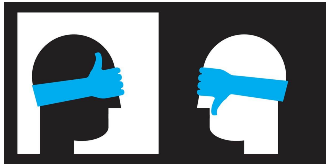
Modularity



Model Compactness



Fairness and Debiasing



Plus Other Crucial Aspects: ***Cross-Cultural Adaptation, Multi-Modal Learning, Commonsense and World Knowledge, User Experience***

UNIVERSITY OF CAMBRIDGE

Language Technology Lab

The talk is largely based on the following papers:

UNIVERSITY OF CAMBRIDGE

Language Technology Lab

Massive thanks to my co-authors!