mSCAN: A Multilingual Benchmark for Compositional Generalisation

Amélie Reymond, Shane Steinert-Threlkeld
UW Linguistics
attr@uw.edu

The SCAN task

SCAN is a classic compositional generalisation benchmark with synthetic data, from Lake and Baroni, 2018

Goal of the task: convert natural language commands to action sequences

Example
Input: jump opposite left and walk thrice
Expected output: LTURN LTURN JUMP WALK WALK WALK

Why make SCAN multilingual?

1. There are multiple compositional generalisation benchmarks… in English
2. Compositional generalization might not work uniformly across languages
3. To evaluate compositional generalization abilities of multilingual LLMs

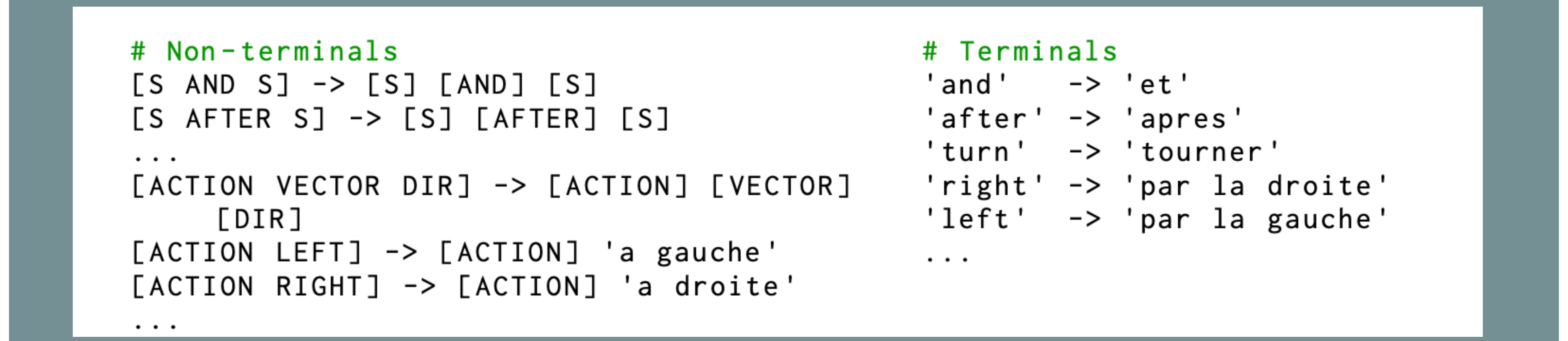
Dataset creation

step 1

Given the original SCAN grammar (Lake and Baroni, 2018), **native speakers** of **French, Mandarin Chinese, Russian and Hindi** provided interpretation functions in their language

step 2

Build transduction rules manually from interpretation functions

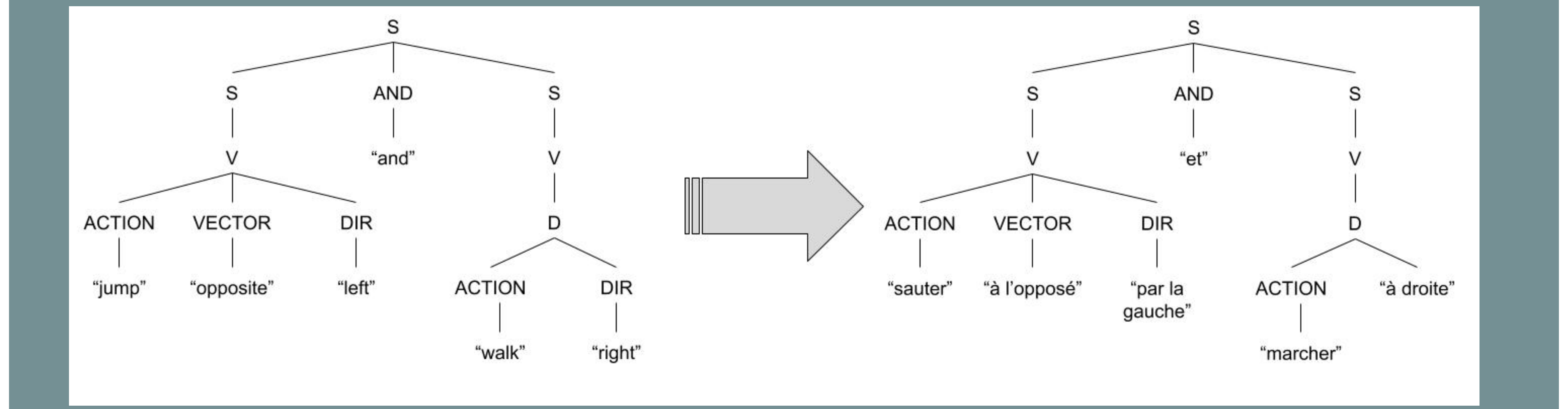


```
# Non-terminals                            # Terminals
[S AND S] -> [S] [AND] [S]                 'and'   -> 'et'
[S AFTER S] -> [S] [AFTER] [S]             'after' -> 'apres'
...                                        'turn'  -> 'tourner'
[ACTION VECTOR DIR] -> [ACTION] [VECTOR]   'right' -> 'par la droite'
     [DIR]                                 'left'  -> 'par la gauche'
[ACTION LEFT] -> [ACTION] 'a gauche'       ...
[ACTION RIGHT] -> [ACTION] 'a droite'
...
```

Example: *English to French transduction rules*

step 3

Use transduction rules to convert English parse trees into target language parse trees

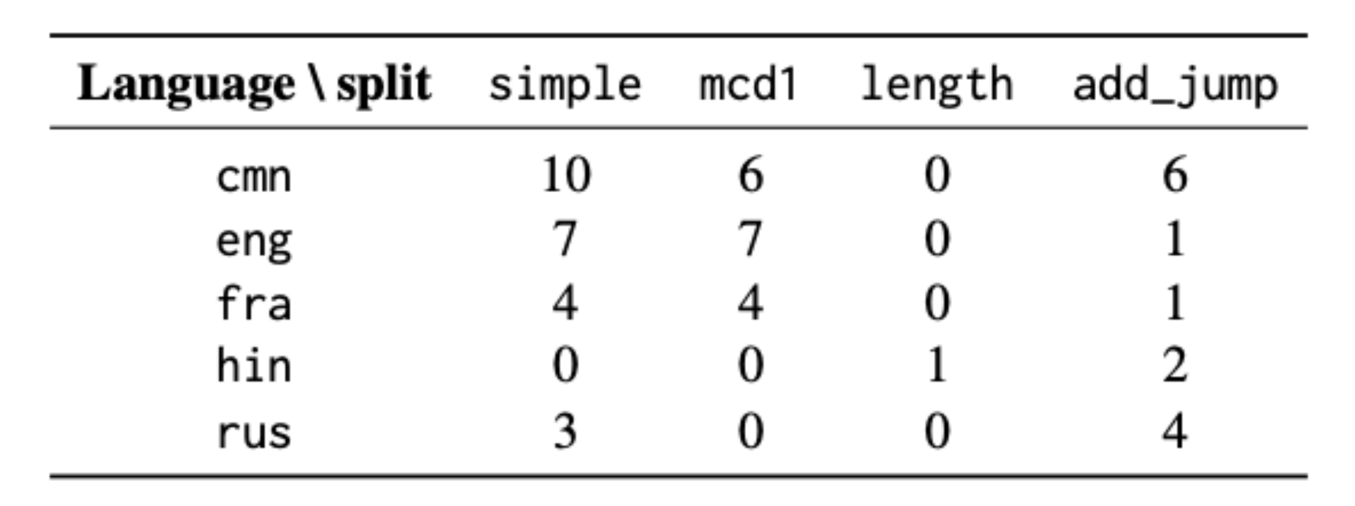


step 4

Serialize parse trees. Re-create original English SCAN splits and maximum compound divergence (MCD) splits (Keysers et al. 2020) in the various languages.

In-context learning experiment

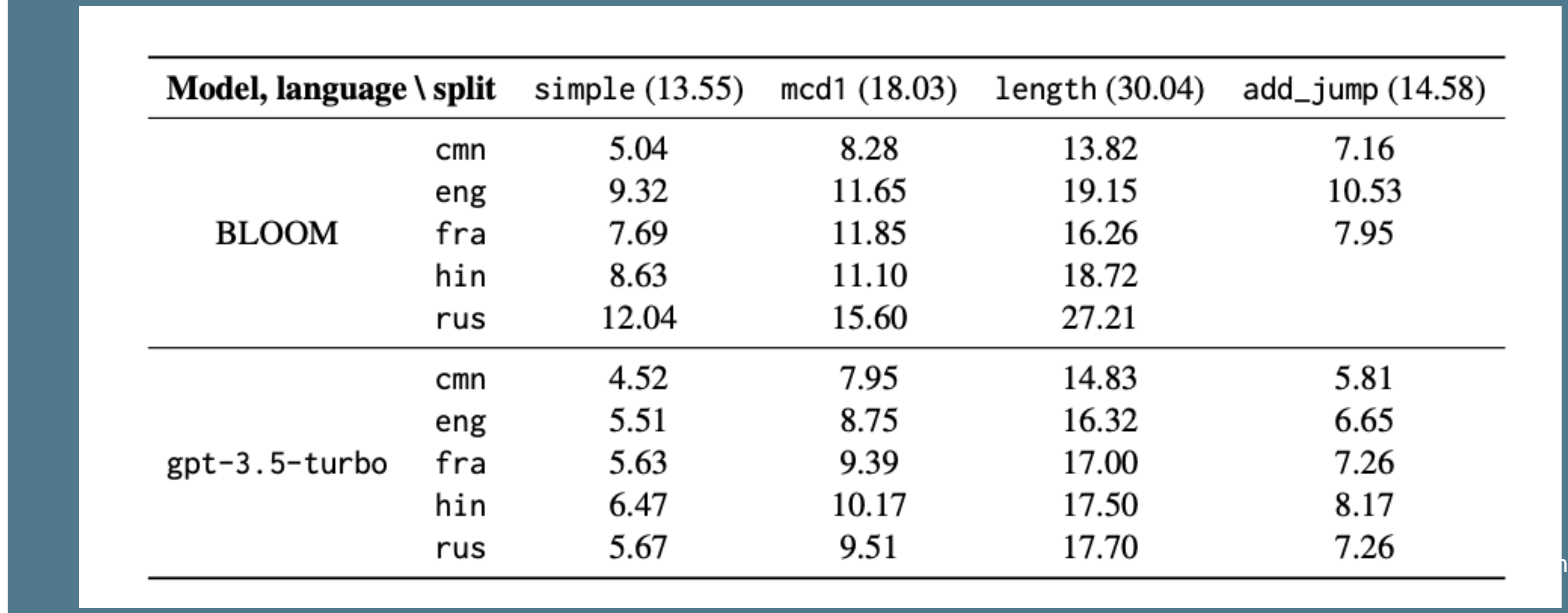
Models: BLOOM and gpt3.5-turbo
Prompt setup: 100 in-context queries, context size of 8 examples



Language \ split	simple	mcd1	length	add_jump
cmn	10	6	0	6
eng	7	7	0	1
fra	4	4	0	1
hin	0	0	1	2
rus	3	0	0	4

exact matches for gpt3.5-turbo: better on Mandarin Chinese (cmn)

Results:
- GPT3.5 got some exact matches
- BLOOM got **none**



Model, language \ split		simple (13.55)	mcd1 (18.03)	length (30.04)	add_jump (14.58)
BLOOM	cmn	5.04	8.28	13.82	7.16
	eng	9.32	11.65	19.15	10.53
	fra	7.69	11.85	16.26	7.95
	hin	8.63	11.10	18.72	
	rus	12.04	15.60	27.21	
gpt-3.5-turbo	cmn	4.52	7.95	14.83	5.81
	eng	5.51	8.75	16.32	6.65
	fra	5.63	9.39	17.00	7.26
	hin	6.47	10.17	17.50	8.17
	rus	5.67	9.51	17.70	7.26

Average edit distance per split. The expected output length is indicated in brackets.

A closer look: edit distance
- Some variation across languages. Surprisingly better results on Mandarin Chinese (cmn) than English (eng).
- Regardless of language, `length` is the most challenging split

Conclusion

- We introduce mSCAN, a multilingual version of SCAN in **French, Mandarin Chinese, Russian and Hindi.**
- It was generated following a **rule-based** procedure, with the consultation of **native speakers**.
- Preliminary experiments show variation across languages, supporting the **importance of multilingual evaluation**.

Links

- Paper: bit.ly/mscan_paper
- Dataset: bit.ly/mscan_data
- Code: bit.ly/mscan_repo