Learner Treebanks and CHILL (Chinese Intelligent Language Learning)

Luis Morgado da Costa
Palacký University Olomouc

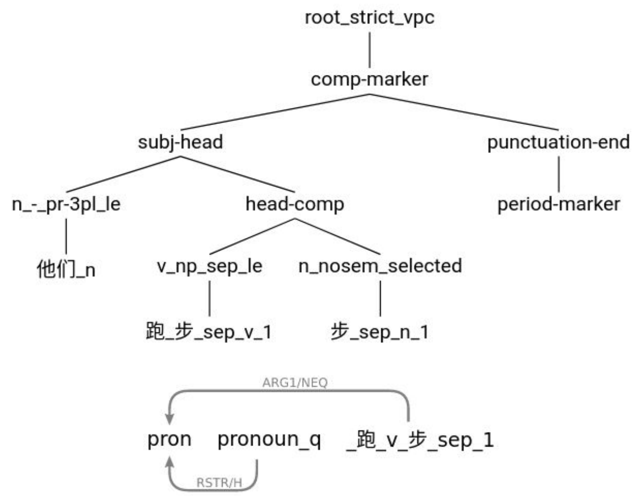
18th July, Fairhaven, US

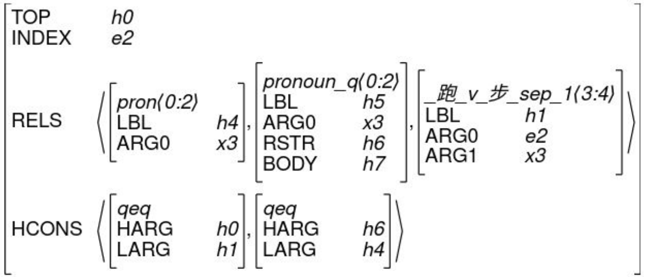
European Commission
Marie Skłodowska-Curie Actions

Univerzita Palackého v Olomouci

ZHONG: A Chinese HPSG Implemented Grammar

- The project started in 2015 (by Fan Zhenzhen), taken up as a small portion of my PhD
- Supposed to be “Meta-Chinese” grammar
- It handles well sentences syntactic structures in low proficiency materials (up to HSK 3)
- Some notable syntactic work includes:
    - 的 constructions (by Zhenzhen)
    - Verbal and adjectival Reduplication
    - Separable verbs (e.g. 生病, 生了病)
    - Aspect (and it’s interactions w/negation)





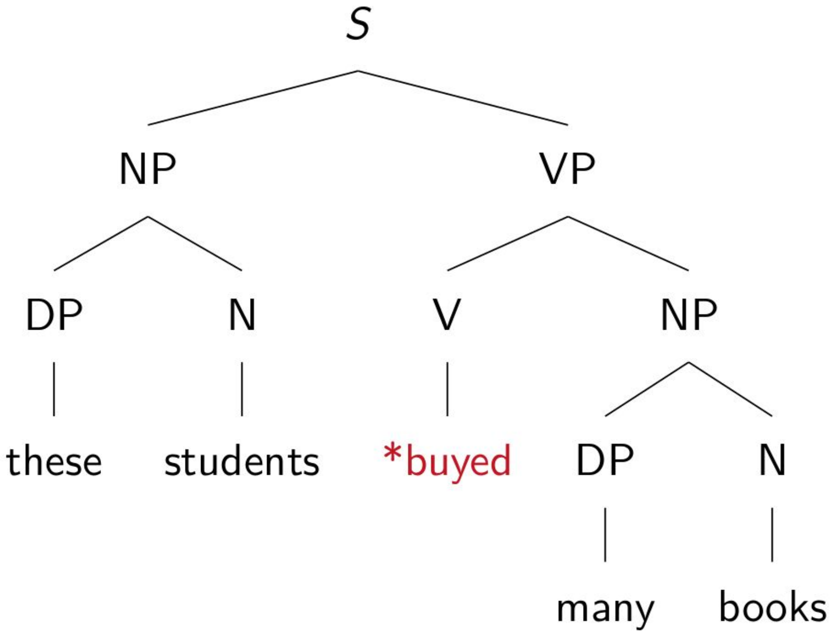
```
TOP     h0
INDEX   e2
RELS    < [pron(0:2)       ], [pronoun_q(0:2) ], [_跑_v_步_sep_1(3:4)] >
          [LBL    h4       ]  [LBL     h5     ]  [LBL     h1         ]
          [ARG0   x3       ]  [ARG0    x3     ]  [ARG0    e2         ]
                              [RSTR    h6     ]  [ARG1    x3         ]
                              [BODY    h7     ]
HCONS   < [qeq    ], [qeq    ] >
          [HARG h0 ]  [HARG h6 ]
          [LARG h1 ]  [LARG h4 ]
```

ZHONG: A Chinese HPSG Implemented Grammar

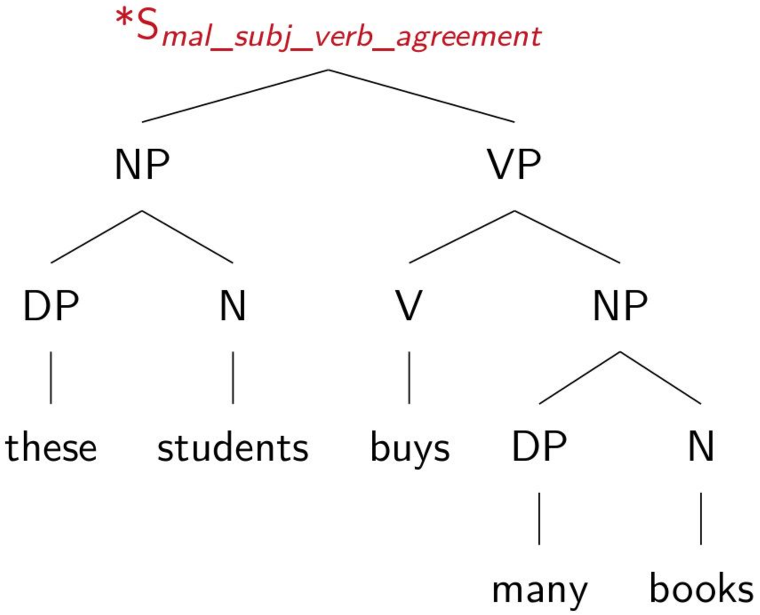
- **MSCA project** – CHILL (Chinese Intelligent Language Learning)
- The grammar should be able to handle up to **HSK 5 at the end of 2023**
- Focus on **NP structure** (quantification, deixis, and cognitive status) **& mal-rules**
- Also In the pipeline (or needing improving):
    - Better treatment of numeric phrase predication
    - Better treatment of passives
    - Comparatives
    - Argument Changing Complements (duration, state, result, potential)

Mal-Rules (Examples)

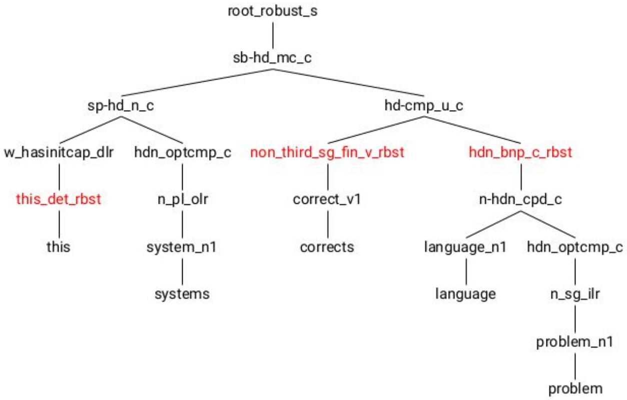
* These students buyed many books.

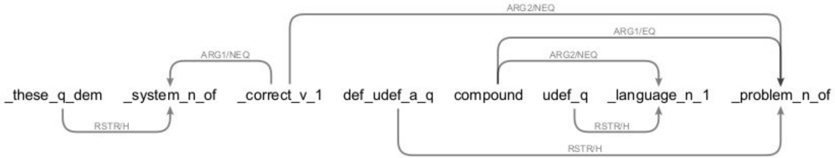


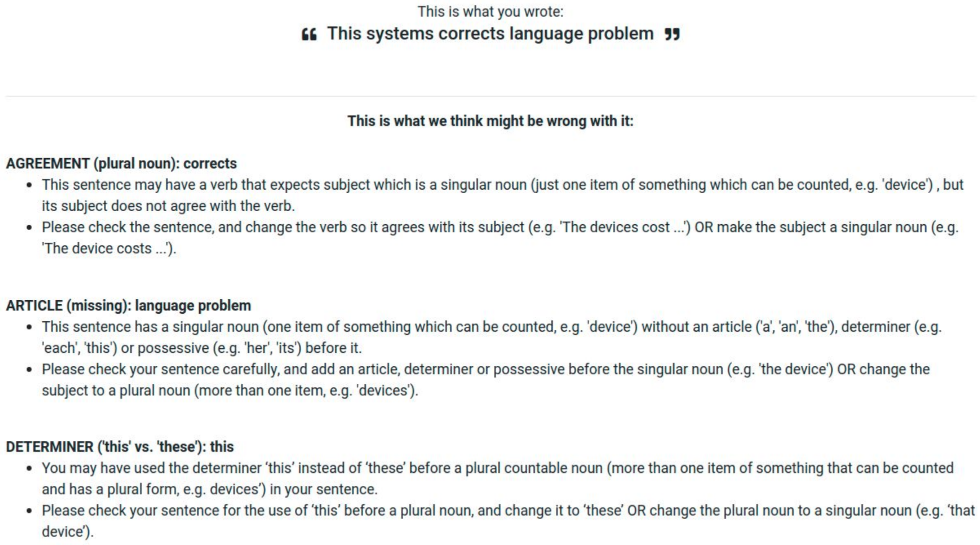
* These students buys many books.



Linking Mal-rules to Corrective Feedback







This is what you wrote:

“ This systems corrects language problem ”

This is what we think might be wrong with it:

AGREEMENT (plural noun): corrects

- This sentence may have a verb that expects subject which is a singular noun (just one item of something which can be counted, e.g. 'device') , but its subject does not agree with the verb.
- Please check the sentence, and change the verb so it agrees with its subject (e.g. 'The devices cost ...') OR make the subject a singular noun (e.g. 'The device costs ...').

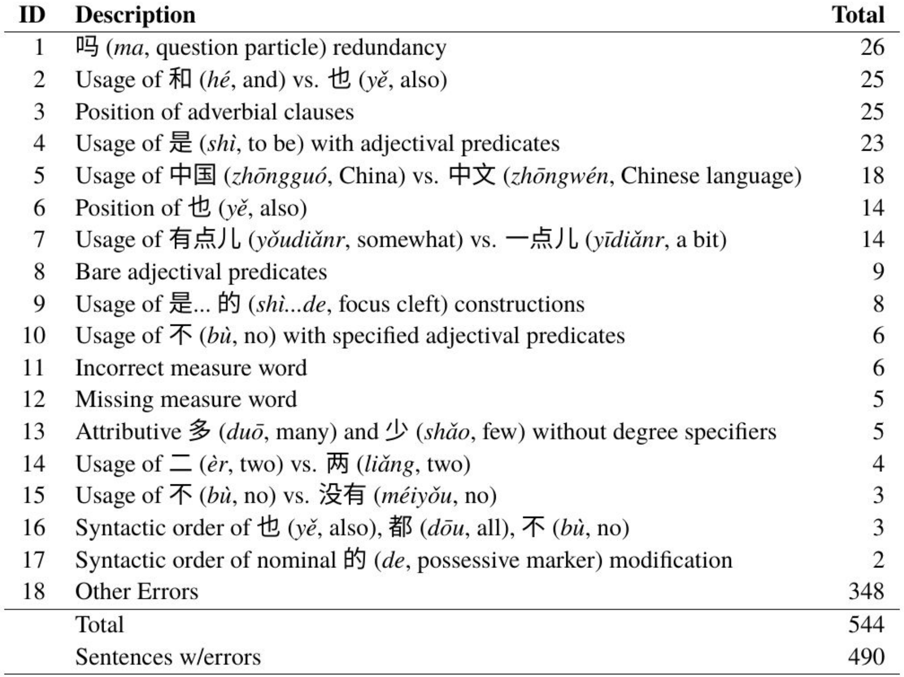
ARTICLE (missing): language problem

- This sentence has a singular noun (one item of something which can be counted, e.g. 'device') without an article ('a', 'an', 'the'), determiner (e.g. 'each', 'this') or possessive (e.g. 'her', 'its') before it.
- Please check your sentence carefully, and add an article, determiner or possessive before the singular noun (e.g. 'the device') OR change the subject to a plural noun (more than one item, e.g. 'devices').

DETERMINER ('this' vs. 'these'): this

- You may have used the determiner ‘this’ instead of ‘these’ before a plural countable noun (more than one item of something that can be counted and has a plural form, e.g. devices’) in your sentence.
- Please check your sentence for the use of ‘this’ before a plural noun, and change it to ‘these’ OR change the plural noun to a singular noun (e.g. ‘that device’).

NTU Corpus of Learner Mandarin (NTUCLM)



ID	Description	Total
1	吗 (*ma*, question particle) redundancy	26
2	Usage of 和 (*hé*, and) vs. 也 (*yě*, also)	25
3	Position of adverbial clauses	25
4	Usage of 是 (*shì*, to be) with adjectival predicates	23
5	Usage of 中国 (*zhōngguó*, China) vs. 中文 (*zhōngwén*, Chinese language)	18
6	Position of 也 (*yě*, also)	14
7	Usage of 有点儿 (*yǒudiǎnr*, somewhat) vs. 一点儿 (*yīdiǎnr*, a bit)	14
8	Bare adjectival predicates	9
9	Usage of 是... 的 (*shì...de*, focus cleft) constructions	8
10	Usage of 不 (*bù*, no) with specified adjectival predicates	6
11	Incorrect measure word	6
12	Missing measure word	5
13	Attributive 多 (*duō*, many) and 少 (*shǎo*, few) without degree specifiers	5
14	Usage of 二 (*èr*, two) vs. 两 (*liǎng*, two)	4
15	Usage of 不 (*bù*, no) vs. 没有 (*méiyǒu*, no)	3
16	Syntactic order of 也 (*yě*, also), 都 (*dōu*, all), 不 (*bù*, no)	3
17	Syntactic order of nominal 的 (*de*, possessive marker) modification	2
18	Other Errors	348
	Total	544
	Sentences w/errors	490

- ≈5,600 sentences (≈2300 after merging repetitions)
- Most error classes were expected
- “Other Errors” included some interesting unexpected classes (e.g. NP predication)
- There is a **long tail of idiosyncratic errors** that are not interesting to name/model
- We are now **collecting data from Czech students** learning Mandarin

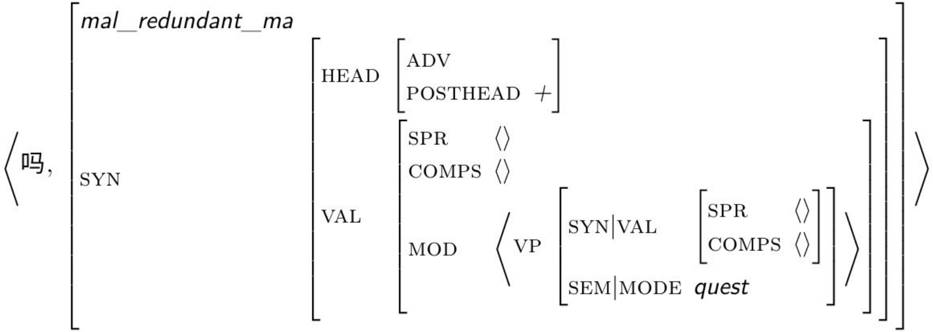
Mal-Rules in ZHONG (吗 Redundancy)

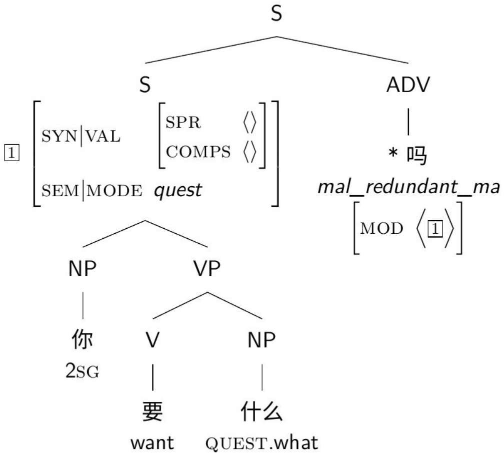
(1) 你 要 什么 ?
2SG want QUEST.what ?
'What do you want?'

(2) *你 要 什么 吗 ?
2SG want QUEST.what QUEST.polar ?
(intended) 'What do you want?'

(3) 你 有 没 有 中文 书 ?
2SG have not have Chinese.language book ?
'Do you have a Chinese textbook?'

(4) *你 有 没 有 中文 书 吗 ?
2SG have not have Chinese.language book QUEST.PART ?
(intended) 'Do you have a Chinese textbook?'



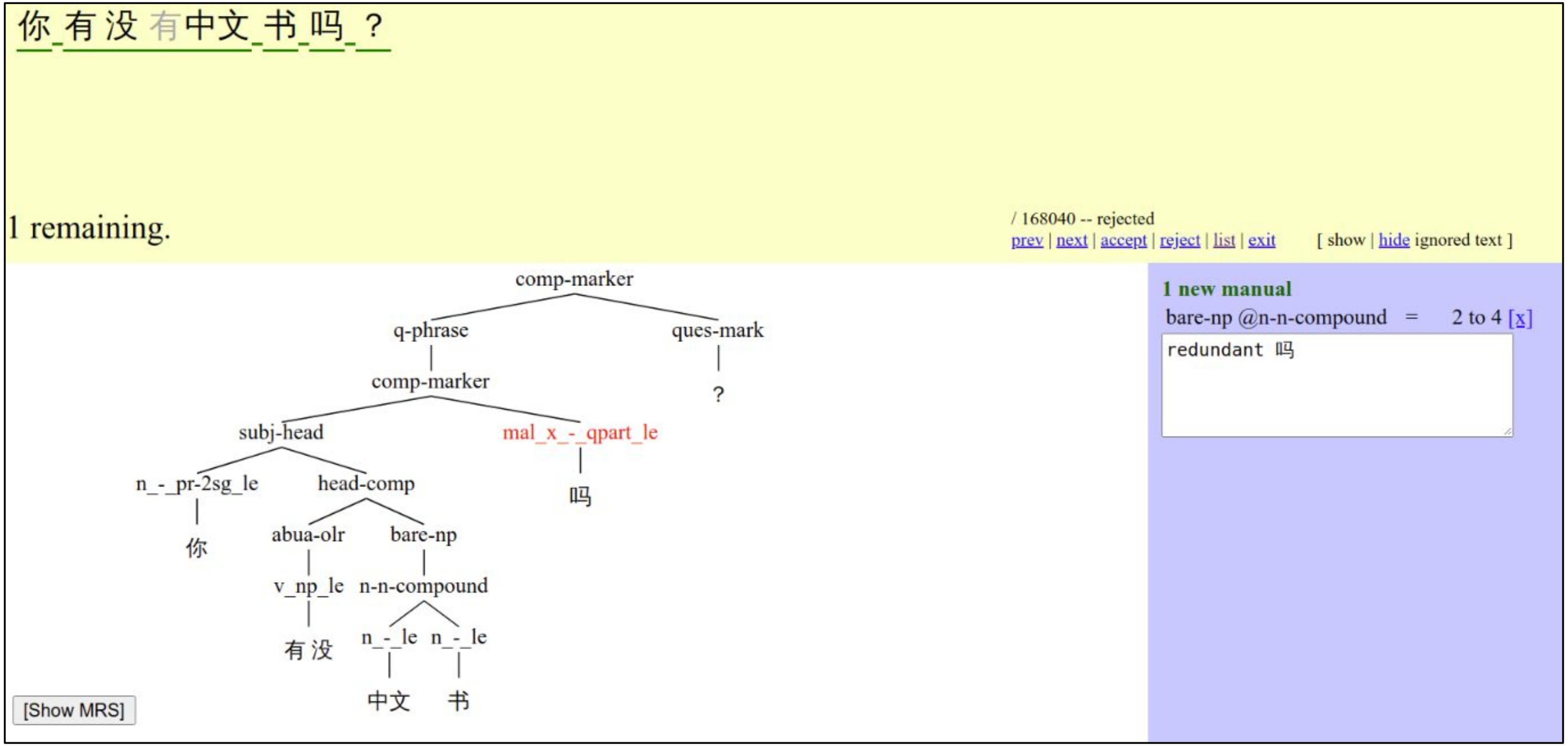


Mal-Rules in ZHONG

- ZHONG now detects more than **60 different mal-rules** (i.e., types of errors)
  - Cover about **50% of the errors** found in the NTUCLM, including:
    - 吗 (ma, question particle) redundancy
    - Clausal coordination with 和 (hé, and)
    - Incorrect position of 也 (yě, also) – e.g., pre-subject
    - 有点儿 (yǒudiǎnr, somewhat) vs. 一点儿 (yīdiǎnr, a bit) confusion
    - Bare NP Predication
    - Missing Measure Words / Classifiers
    - 不 (bù, no) vs. 没有 (méiyǒu, no) confusion
    - 二 (èr, two) vs. 两 (liǎng, two) confusion
    - **Misspellings** (Not sure if they should be handled by the grammar)
    - etc.
- Corrective feedback messages and web-app (for classrooms) is *in progress*

Grammar / Mal-rules Demo: https://www.luismc.com/itell/delphin_analyser

The Mandarin Learner Treebank



你_有 没 有中文_书_吗_？

1 remaining.

/ 168040 -- rejected

prev | next | accept | reject | list | exit [show | hide ignored text]

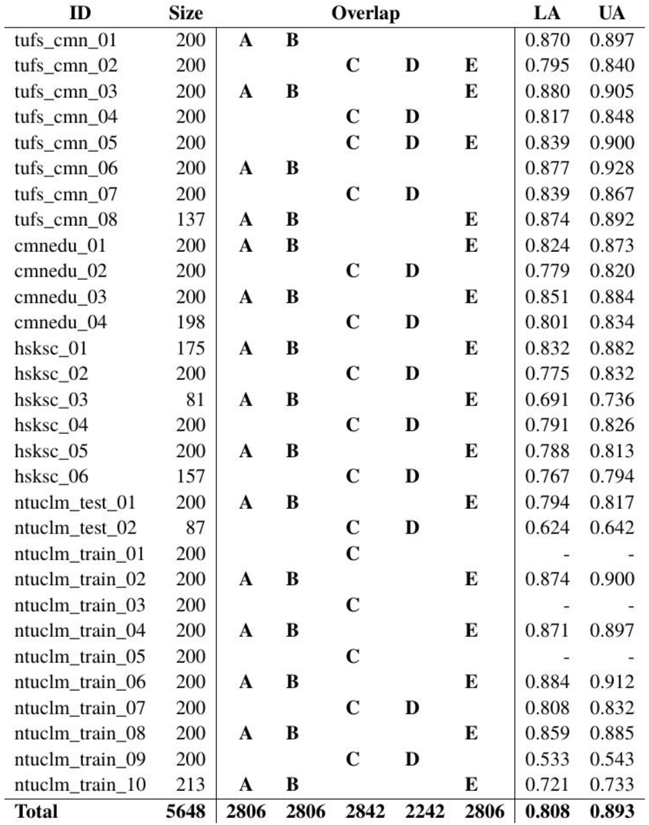
1 new manual

bare-np @n-n-compound = 2 to 4 [x]

```
redundant 吗
```

The Mandarin Learner Treebank

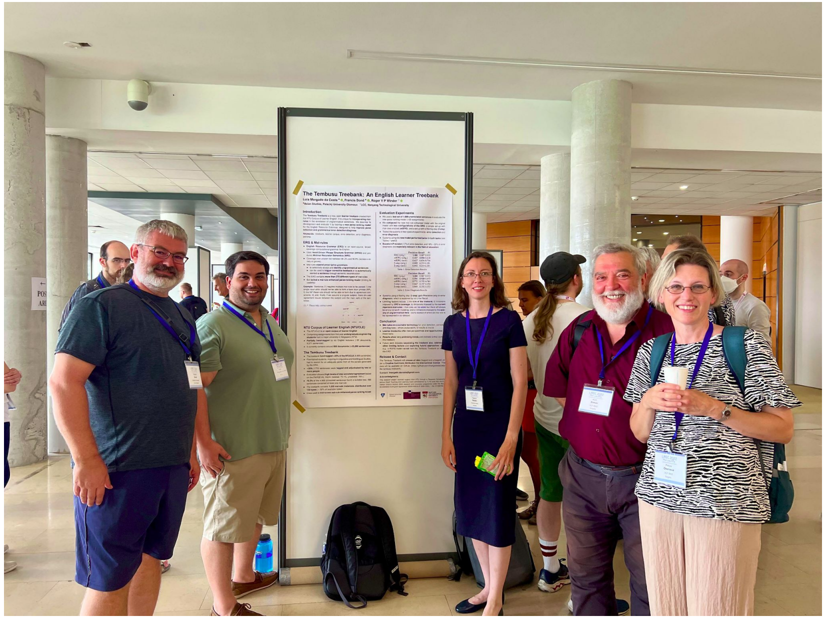
- Treebanked over 5600 sentences manually
- 5 trained student assistants (w/overlap)
- Includes **textbook and learner data**
- Trained a new parse-ranking model
- Improved Grammatical <u>Error Detection</u>
  - 88% Precision (top-parse), 41% Recall
- Improved Grammatical <u>Error Diagnosis</u>
  - 89% Precision (top-parse), 47% Recall
- Moving into Tatoeba



ID	Size	Overlap					LA	UA
tufs_cmn_01	200	A	B				0.870	0.897
tufs_cmn_02	200			C	D	E	0.795	0.840
tufs_cmn_03	200	A	B			E	0.880	0.905
tufs_cmn_04	200			C	D		0.817	0.848
tufs_cmn_05	200			C	D	E	0.839	0.900
tufs_cmn_06	200	A	B				0.877	0.928
tufs_cmn_07	200			C	D		0.839	0.867
tufs_cmn_08	137	A	B			E	0.874	0.892
cmnedu_01	200	A	B			E	0.824	0.873
cmnedu_02	200			C	D		0.779	0.820
cmnedu_03	200	A	B			E	0.851	0.884
cmnedu_04	198			C	D		0.801	0.834
hsksc_01	175	A	B			E	0.832	0.882
hsksc_02	200			C	D		0.775	0.832
hsksc_03	81	A	B			E	0.691	0.736
hsksc_04	200			C	D		0.791	0.826
hsksc_05	200	A	B			E	0.788	0.813
hsksc_06	157			C	D		0.767	0.794
ntuclm_test_01	200	A	B			E	0.794	0.817
ntuclm_test_02	87			C	D		0.624	0.642
ntuclm_train_01	200			C			-	-
ntuclm_train_02	200	A	B			E	0.874	0.900
ntuclm_train_03	200			C			-	-
ntuclm_train_04	200	A	B			E	0.871	0.897
ntuclm_train_05	200			C			-	-
ntuclm_train_06	200	A	B			E	0.884	0.912
ntuclm_train_07	200			C	D		0.808	0.832
ntuclm_train_08	200	A	B			E	0.859	0.885
ntuclm_train_09	200			C	D		0.533	0.543
ntuclm_train_10	213	A	B			E	0.721	0.733
Total	**5648**	**2806**	**2806**	**2842**	**2242**	**2806**	**0.808**	**0.893**

By the way... The Tembusu Treebank is here!

Morgado da Costa, Luis and Bond, Francis and Winder, Roger V. P. (2022). The **Tembusu Treebank: An English Learner Treebank**. *Proceedings of the 13th Conference on Language Resources and Evaluation*. European Language Resources Association. Marseille, France.



Some Challenges Lying Ahead

Some Current Challenges

- **Integrate Segmentation**
  - Integrate external segmenters / POS-taggers? (unknown word handling)
  - Character/pinyin-based parsing (I need some help with REPP)
- **Lexicon Management**
  - Tools to keep results of lexical tests and generate lexicon
  - Possibility of linking and or merging with the Chinese Open Wordnet
- **Treebanks / Release Cycle:**
  - Building, Formatting and Sharing Treebanks (SIG?), incl. tools (LTDB?)
- **Data Collection:**
  - Streamline learner data collection through some apps
- **End the “meta-chinese” approach:**
  - out-of-date, difficult to manageable, not aligned with current goals