L1-L2 Parallel Dependency Treebank as Learner Corpus

John Lee, Keying Li, Herman Leung

City University of Hong Kong

香港城市大學
City University of Hong Kong

Outline

- **Introduction**
- Parallel treebanks
- Learner corpora
- L1-L2 parallel treebank as learner corpus
  - Case study

香港城市大學
City University of Hong Kong

Introduction

- A learner corpus consists of text written by language learners
  - Typically indicates learner errors with:
    - Error tags
    - Target hypothesis

He <MV> null | is </MV> happy.

Error tag:
M(issing) V(erb)

Target hypothesis:
Corrected version of sentence

香港城市大學
City University of Hong Kong

Introduction

- Learner corpora facilitate retrieval of large number of samples for quantitative studies
  - Error Analysis
    - What are the most common error categories in learner text?
  - Contrastive Interlanguage Analysis
    - What words or structures are overused or underused by learners, compared to native speakers?

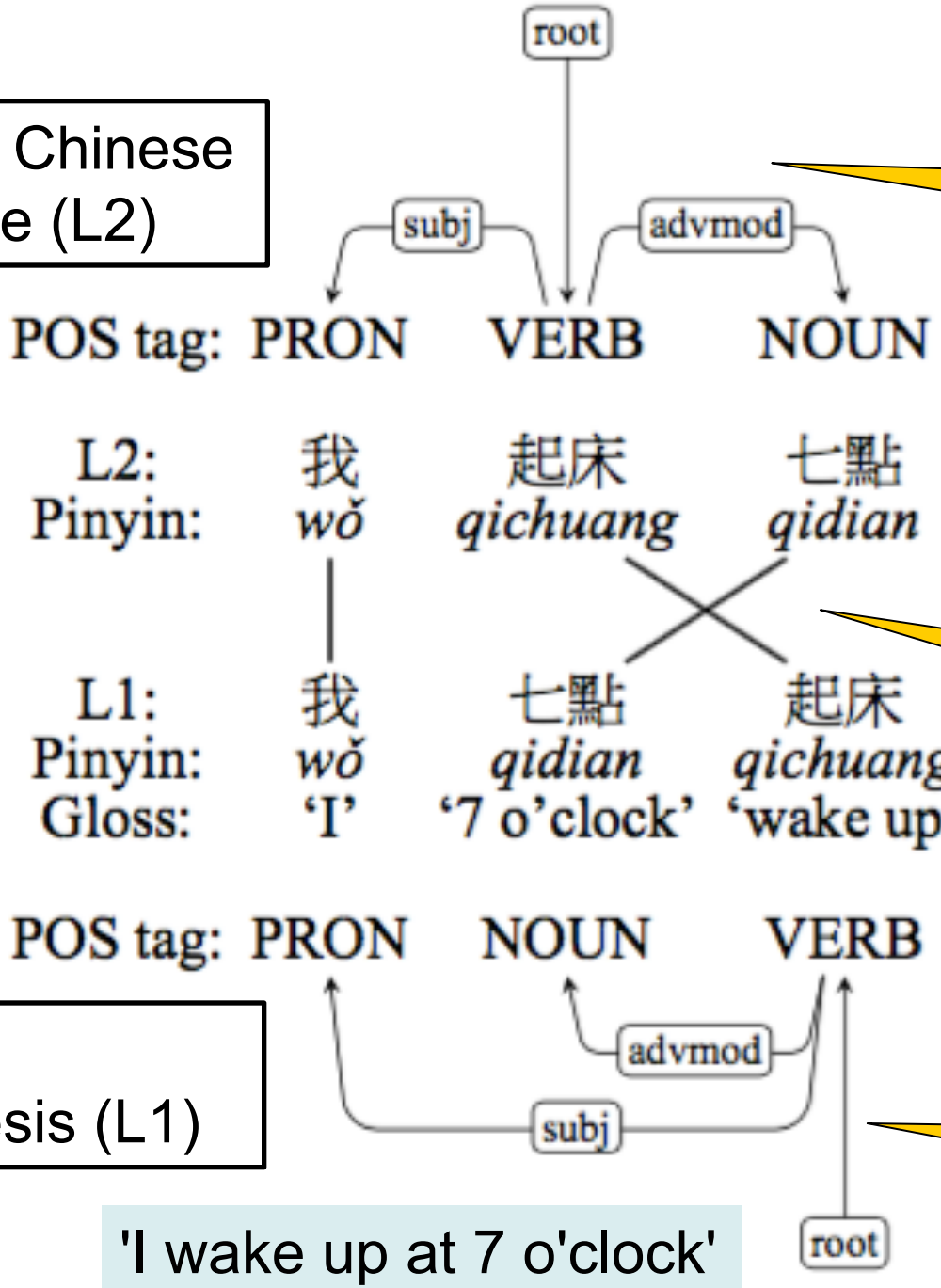
香港城市大學
City University of Hong Kong

Introduction

- We propose annotating a learner corpus as an *L1-L2 parallel treebank*
  - L2 treebank
    - Learner sentences, with syntactic trees
  - L1 treebank
    - Target hypotheses, with syntactic trees
  - Word alignment between L1 and L2 trees

香港城市大學
City University of Hong Kong

Learner Chinese sentence (L2)



Target hypothesis (L1)

'I wake up at 7 o'clock'

Syntactic tree for L1

POS tag for L1

Word alignment

POS tag for L2

Syntactic tree for L2

香港城市大學
City University of Hong Kong

Introduction

- This paper discusses:
  - Advantages of using a parallel L1-L2 treebank to analyze learner language
    - More flexible retrieval of different error types
  - Case study on word-order errors
    - Evaluation on accuracy in retrieving different types of word-order errors
    - Based on a small parallel Chinese L1-L2 treebank

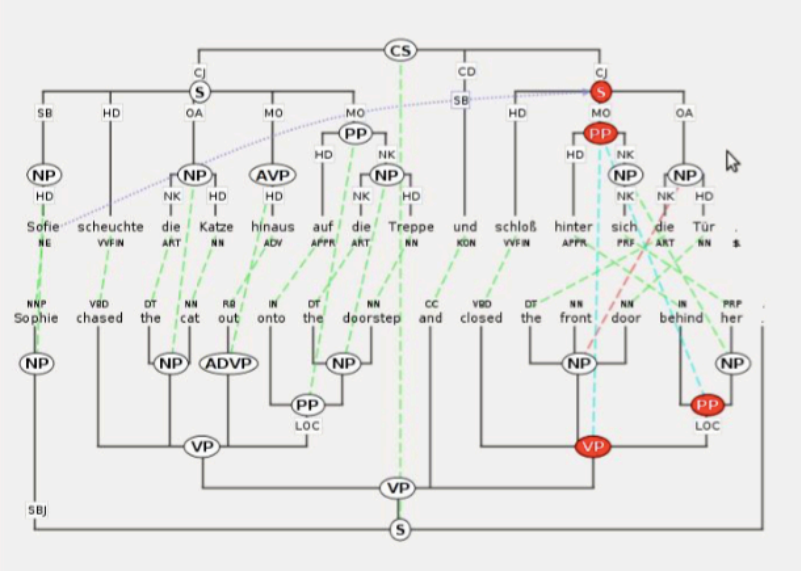
香港城市大學
City University of Hong Kong

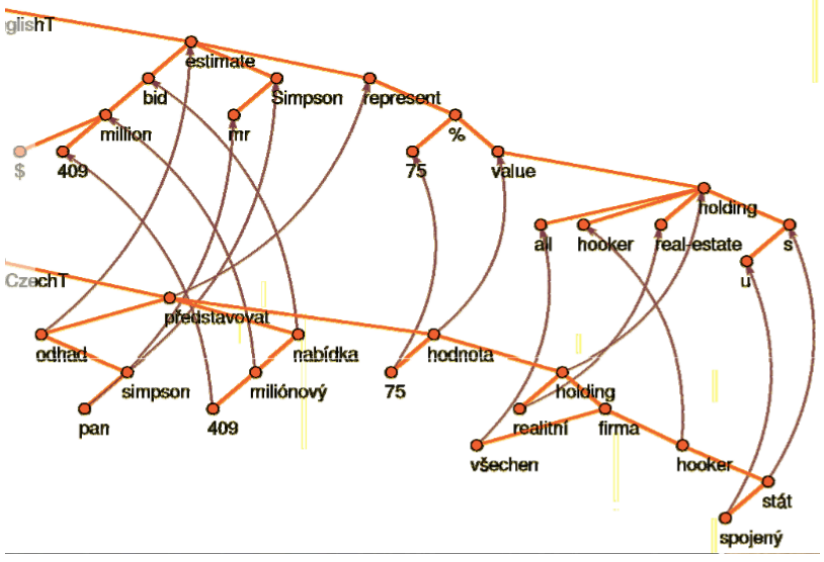
Outline

- Introduction
- **Parallel Treebanks**
- Learner Corpora
- L1-L2 parallel treebank as learner corpus
  - Case study

Parallel treebanks

- Parallel treebanks increasingly available
  - Czech-English, English-French, English-German, English-German-Swedish, English-Swedish-Turkish (Cmejrek et al. 2003; Hansen-Schirra et al., 2006; Ahrenberg, 2007; Hearne and Way, 2006, Megyesi et al., 2010)





(Cmejrek et al., 2003; Volk & Marek, 2011)

香港城市大學
City University of Hong Kong

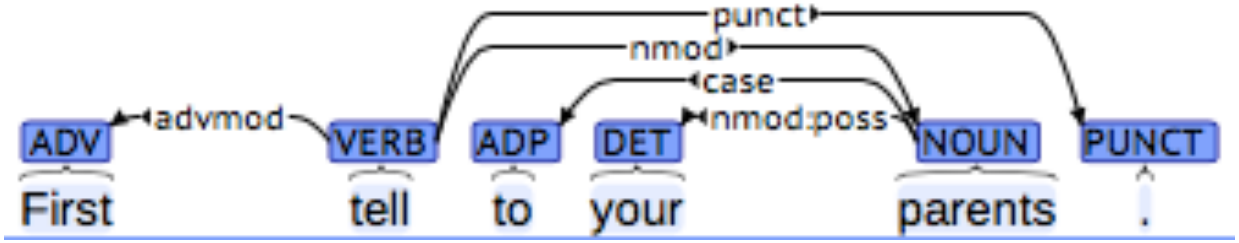
Parallel treebanks

- Parallel treebanks support quantitative comparison between languages
  - Translation correspondence
  - Typological features
    - Copula construction, predicate structure, etc. (Sulger et al., 2013)

- An *L1-L2 parallel treebank* can similarly support comparison between a language and an interlanguage

Parallel treebanks

- Treebanks have been constructed for learner English
  - Dependency treebanks (Berzak et al., 2016; Ragheb and Dickinson, 2014)
  - Constituent treebanks (Nagata and Sakaguchi, 2016)
  - Not yet any L1-L2 parallel treebank



[Berzak et al., 2016]

香港城市大學
City University of Hong Kong

Outline

- Introduction
- Parallel Treebanks
- **Learner Corpora**
- L1-L2 parallel treebank as learner corpus
  - Case study

香港城市大學
City University of Hong Kong

Error tags

- NUCLE error tagset (Dahlmeier et al., 2013)

Verb tense	Noun number
Verb modal	Noun possessive
Missing verb	Pronoun form
Verb form	Pronoun reference
Subject-verb agreement	Wrong collocation
Article or determiner	Acronyms
Runons	Word form
Dangling modifiers	Tone
Parallelism	Subordinate clause
Fragment	

香港城市大學
City University of Hong Kong

Error tags

Word-level:
action verb (*v*), auxiliary (*aux*), stative verb (*vs*), noun (*n*), pronoun (*pron*), conjunction (*conj*), preposition (*p*), numeral (*num*), demonstrative (*det*), measure word (*cl*), sentential particle (*sp*), aspectual particle (*asp*), adverb (*adv*), structural particle (*de*), question word (*que*), plural suffix (*plural*)

Grammatical Function-level:
subject (*sub*), object (*obj*), noun phrase (*np*), verb phrase (vp), preposition phrase (*pp*), modifier (*mod*), time expression (*time*), place expression (*loc*), transitivity (*tran*), separable structure (*vo*), [numeral /determiner+measure] phrase (*dm*),

---

Sentence Pattern-level:
complex noun clause (*rel*), 把 (*ba*), 被 (*bei*), 讓 (*rang*), 是 (*shi*), 有 (*you*), other patterns (*pattern*)

Mixture:
formation (*form*), ambiguity of syntactic or meaning (*sentence*)

- Test of Chinese as a Foreign Language Learner corpus (Lee et al., 2016)

香港城市大學
City University of Hong Kong

Limitations

- Error tags impose a fixed error typology

- Limited corpus re-use
  - Difficult to develop a robust and general-purpose error typology
  - Cannot cover “all” error categories of potential interest
  - Researchers need to re-annotate for their own studies

香港城市大學
City University of Hong Kong

Limitations

- Limited corpus interoperability
  - Granularity of error tagset varies among corpora
    - E.g., Learner English: NUCLE (27 tags) vs NICT Japanese Learner English Corpus (46 tags) vs Cambridge Learner Corpus (80 tags)
  - To leverage multiple corpora, one would need to map error categories from one corpus to another
    - Difficult because of differences in definition

香港城市大學
City University of Hong Kong

Outline

- Introduction
- Parallel Treebanks
- Learner Corpora
- **L1-L2 parallel treebank as learner corpus**
  - Case study

Tree search for error retrieval

- Many error categories can be expressed as a search query on POS tags

L2	Furniture	look	good
POS tag	NN	VB	JJ

L1	Furniture	looks	good
POS tag	NN	VBZ	JJ

Search on aligned VB-VBZ words can retrieve subject-verb agreement errors

Tree search for error retrieval

- But POS tags alone are often not sufficient
  - E.g., change in POS might be a consequence of other errors

L2	Furnitures	look	good
POS tag	NNS	VB	JJ

L1	Furniture	looks	good
POS tag	NN	VBZ	JJ

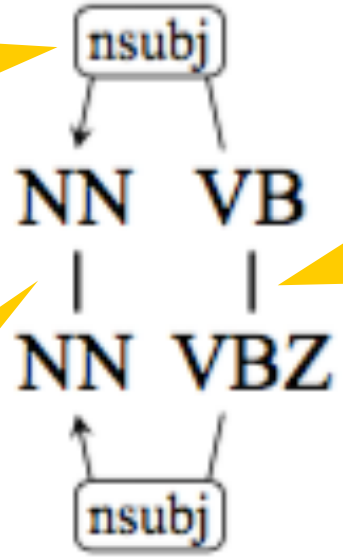
Not a subject-verb agreement error, but a noun number error

Tree search for error retrieval

- More precise search is possible with dependency relations

Both verbs have the same noun subject

The noun subject is not changed



Verb changed from base form (VB) to present third-person singular (VBZ)

香港城市大學
City University of Hong Kong

Outline

- Introduction
- Parallel Treebanks
- Learner Corpora
- L1-L2 parallel treebank
  - **Case study**

Chinese word-order errors

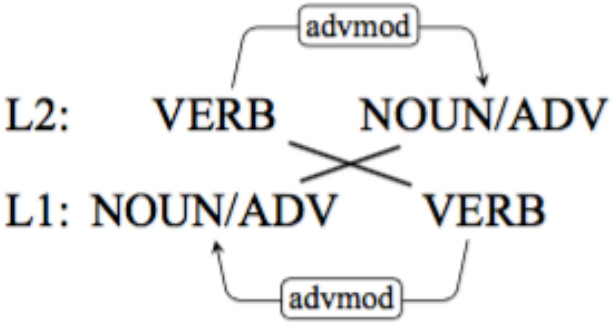
- Types of Chinese word-order errors
  - 3 categories proposed by Ko (1997)
    - Time/Place Words
    - Modification Structures
    - Topic-comment Relations
  - 27 categories proposed by Jiang (2009)
  - Current Chinese learner corpora do not provide this granularity
    - Impossible to distinguish between these categories

香港城市大學
City University of Hong Kong

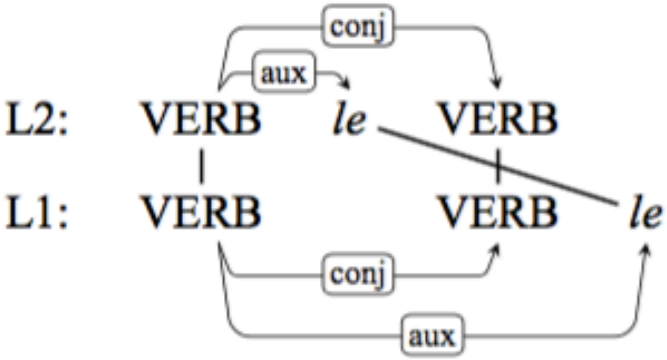
Data

- Dev set: 58 sentence pairs from Jiang (2009)
  - Manually developed 30 parse tree patterns for 10 error categories
  - Annotated sentence with Universal Dependencies
    - Based on scheme proposed by Lee et al. (2017)
- Test set: 114 sentences

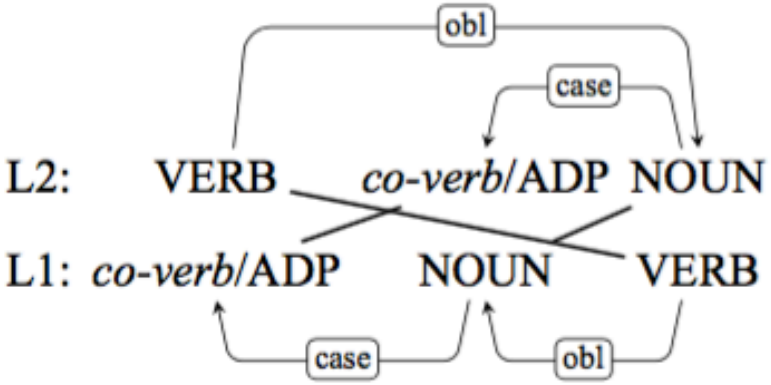
(a) Modifiers + V (Adverb + V)
L2: 我去第一次中國...
wo qu/VERB *diyici*/NOUN *zhongguo*
'I' 'go' 'first time' 'China'
L1: 我第一次去中國...
wo diyici/NOUN *qu*/VERB *zhongguo*
'I' 'first time' 'go' 'China'
"I go for the first time to China ..."



(b) Action Series (LE position)
L2: 我們去了參觀故宮
women qu/VERB *le canguan*/VERB *gugong*
'we' 'go' LE 'visit' 'Forbidden City'
L1: 我們去參觀了故宮
women qu/VERB *canguan*/VERB *le gugong*
'we' 'go' 'visit' LE 'Forbidden City'
"We went to visit the Forbidden City"



(c) Locative Expressions (Location + V)
L2: 你做什麼在這裡
ni zuo/VERB *shenme zai*/ADP *zheli*/NOUN
'you' 'do' 'what' 'at' 'here'
L1: 你在這裡做什麼
ni zai/ADP *zheli*/NOUN *zuo*/VERB *shenme*
'you' 'at' 'here' 'do' 'what'
"What are you doing here?"



香港城市大學
City University of Hong Kong

Results

Error type	Precision	Recall
Time expressions	0.92	0.92
Modifiers + V	0.50	0.50
Action Series	0.65	0.85
Locative expressions	0.91	0.77
Subsidiary Relations	1.00	0.80
Beneficiary	1.00	0.56
Modifiers + N	0.89	1.00
DE position	1.00	0.38
Topic-comment	0.83	0.71
Question	1.00	0.50

香港城市大學
City University of Hong Kong

Conclusion

- An L1-L2 parallel treebank offers some advantages as learner corpus
  - Corpus re-use
  - Corpus interoperability
- A case study on Chinese word-order errors demonstrates its potential