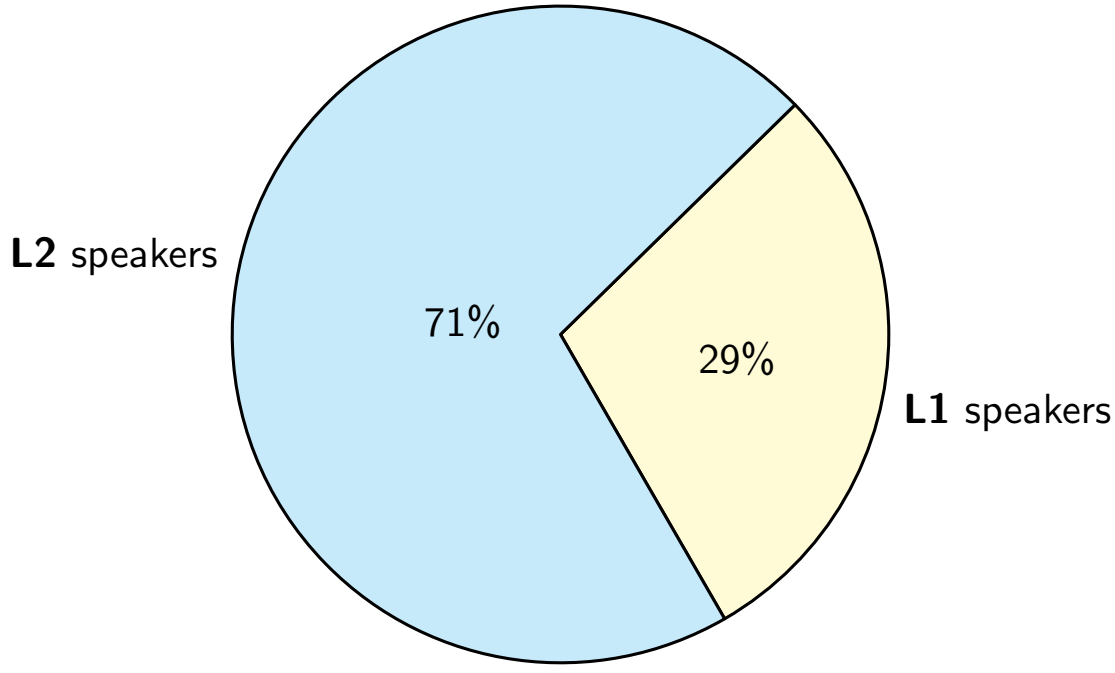
Lecture 16: Semantic Processing for Code-Mixed Languages

L98: Introduction to Computational Semantics

Weiwei Sun

Department of Computer Science and Technology
University of Cambridge

Michaelmas 2024/25



from Ethnologue (2019, 23rd edition);

898.4 million ESL speakers!

Lecture 16: Semantic Processing for Code-Mixed Languages

1. Code-mixing

Code Mixing

Code mixing/switching

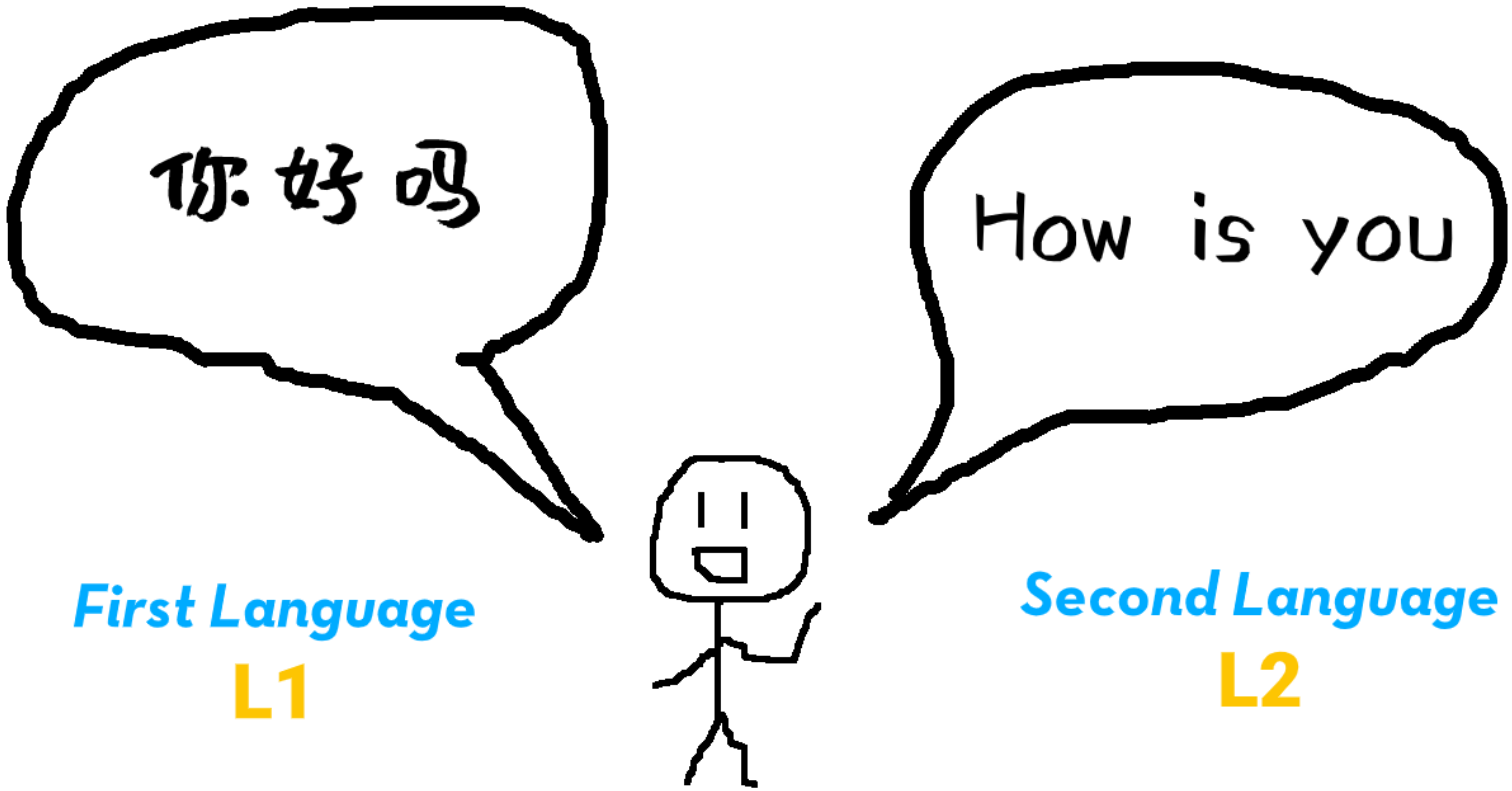
Code-mixing/switching: a speaker alternates between two or more languages in the context of a single conversation or situation.

Code-switching in Hong Kong

The English word "sure" / "cute" is mixed into an otherwise Cantonese sentence.

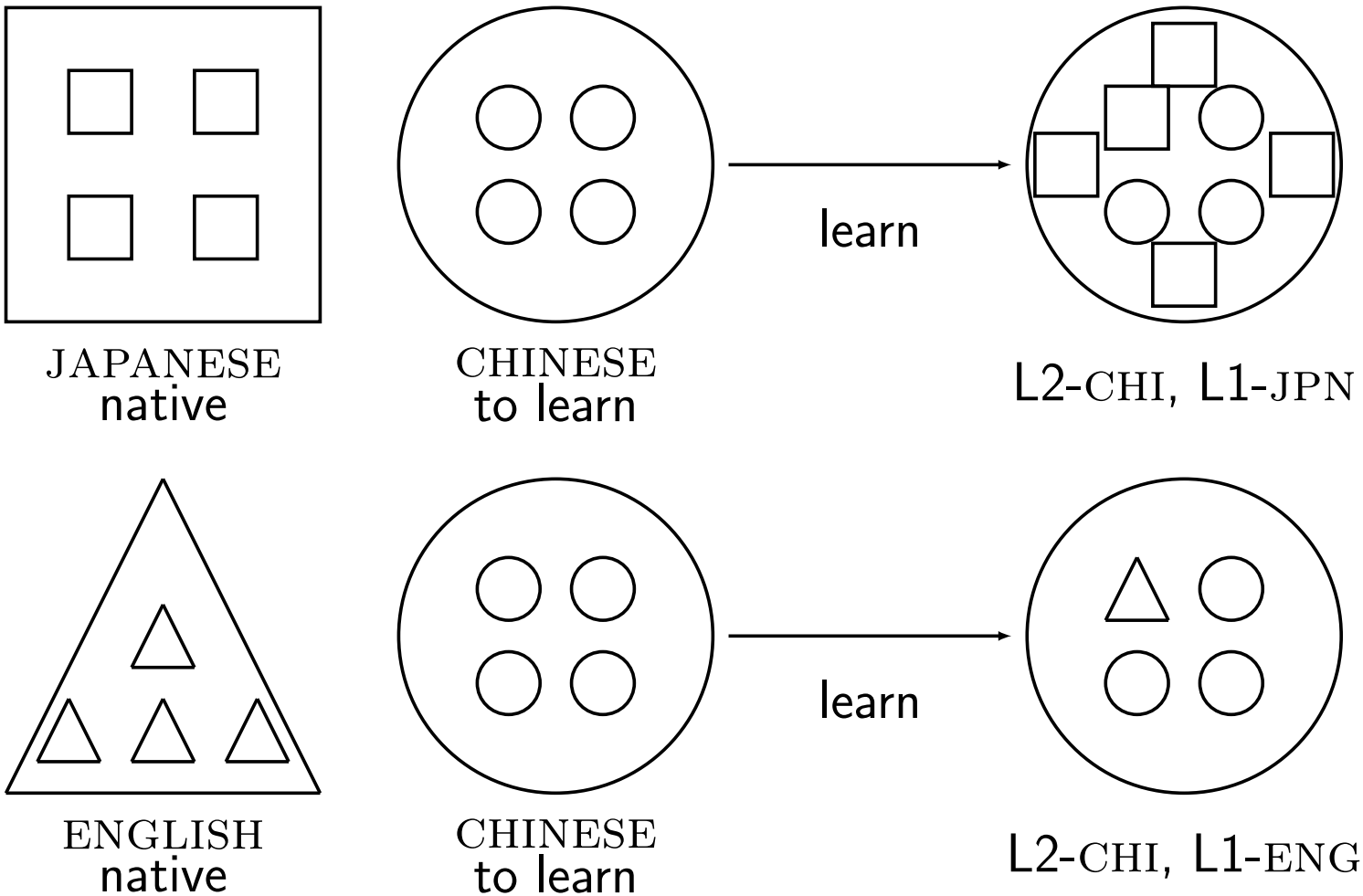
- 我唔sure
- cu唔cute啊

First languages, second languages, cross-lingual transfer



L1 has an influence on L2

Something like



Language acquisition

- Frist language acquisition
- Second language acquisition
- Monolingual FLA
- Bilingual FLA
- Syntactic acquisition
- Semantic acquisition

Universals

Noun Phrase Accessibility Hierarchy

Subject $\succ$ direct object $\succ$ indirect object $\succ$ oblique $\succ$ genitive $\succ$ object of comparison

If a language can relativize on a position on the hierarchy, then any other higher position can also be relativized on.

(1) a. the man who I am taller than $\triangleright$ object of comparison

b. the man whose father I know $\triangleright$ genitive

For example, if a language allows (1a), then it allows (1b).

A universal of SLA

L2 learners find relative clauses higher on the hierarchy easier to acquire.

Annotating English as a Second Language

Annotating second language data

There is naturally a need to automatically annotate second language data with rich lexical, syntactic, semantic and even pragmatic information.

Annotating second language data

There is naturally a need to automatically annotate second language data with rich lexical, syntactic, semantic and even pragmatic information.

High-performance automatic annotation,

- from an engineering perspective, enables deriving high-quality information by structuring this specific type of data, and
- from a scientific perspective, enables quantitative studies for Second Language Acquisition, which is complementary to hands-on experiences in interpreting second language phenomena.

Data: Reddit (https://www.reddit.com)

Large-scale L2 texts are available!

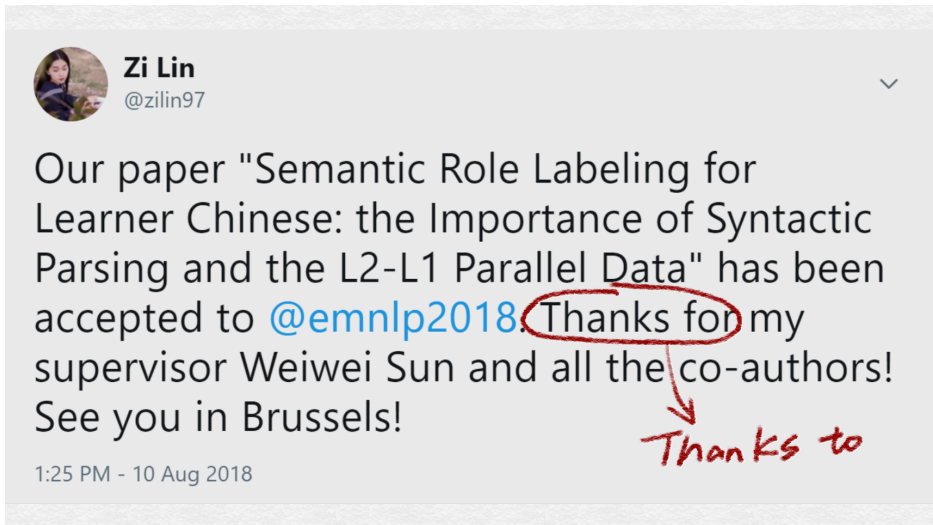
250M native and non-native English sentences (3.8B tokens), covering over 45K authors from 50 countries [**?**]

L1	Sentence
French	*I have to go to the Dr. to do a* **rapid** *check on my heart stability.*
French	*Maybe put every name through a manual* **approbation** *pipeline so it ensures quality.*
French	*Polls have shown public* **approbation** *for this law is somewhere between 58% and 65%, and it has been a strong promise during the presidential campaign.*
Italian	*The event was even more shocking because the* **precedent** *evening he wasn't sick at all.*

Learner texts are everywhere . . .

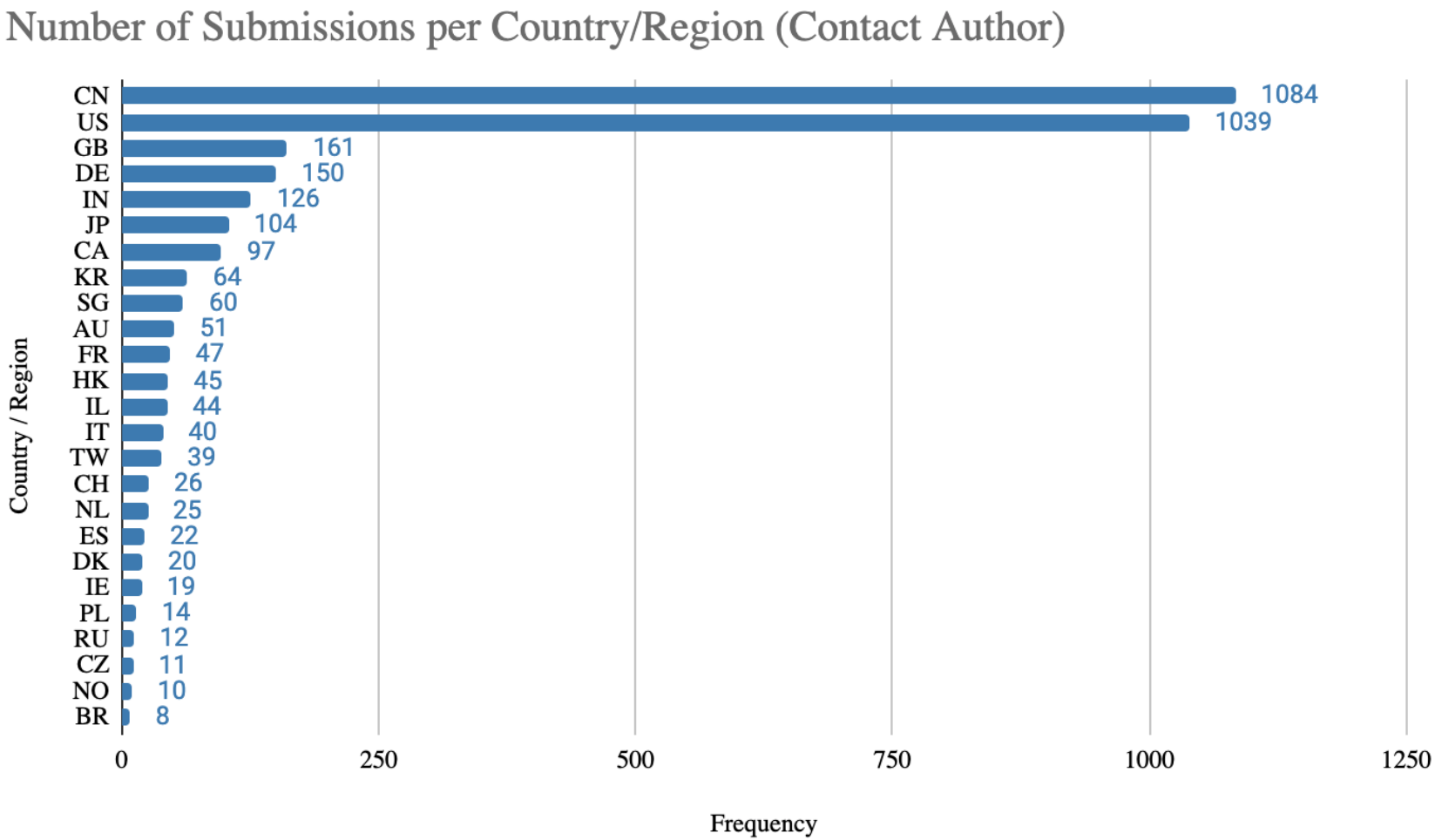
Language Tests

Learner texts are everywhere . . .



Social Network

Learner texts are everywhere . . .



https://acl2020.org/blog/general-conference-statistics/

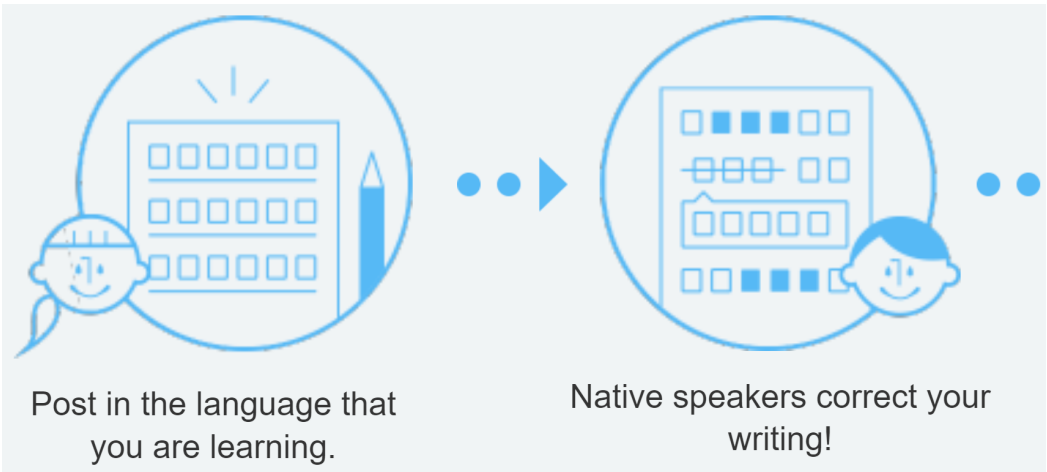
and perhaps your papers

Data: LANG-8 (http://lang-8.com)

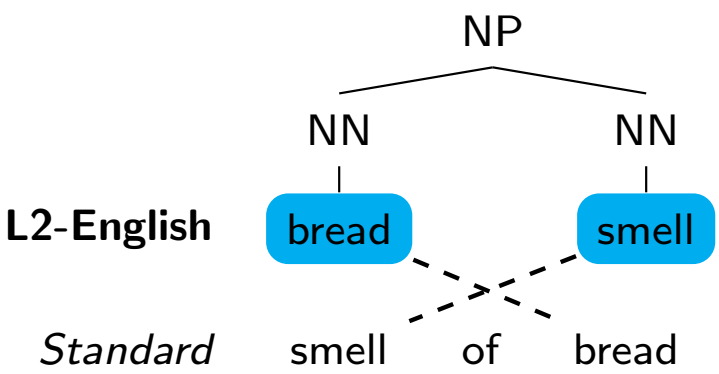
Large-scale L2-L1 parallel texts are available!

6.8M English sentence pairs and 720K Chinese sentence pairs.

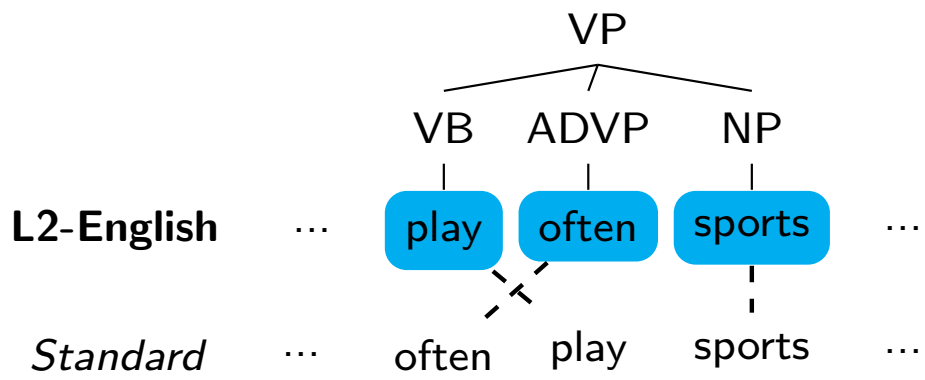
L2 speaker	城市里的人能度过多方面的生活。
corrected	城市里的人能过丰富多彩的生活。
L2 speaker	You know what should I done.
corrected	You know what I should have done.



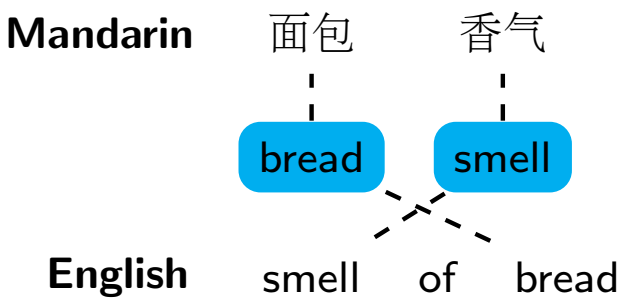
Patterns of cross-lingual transfer

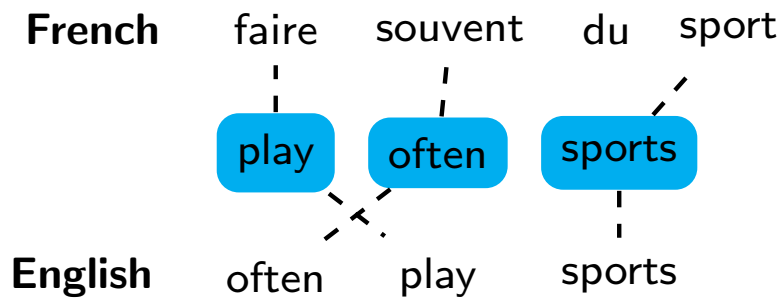


NP(x0:NN x1:NN) → x1 x0

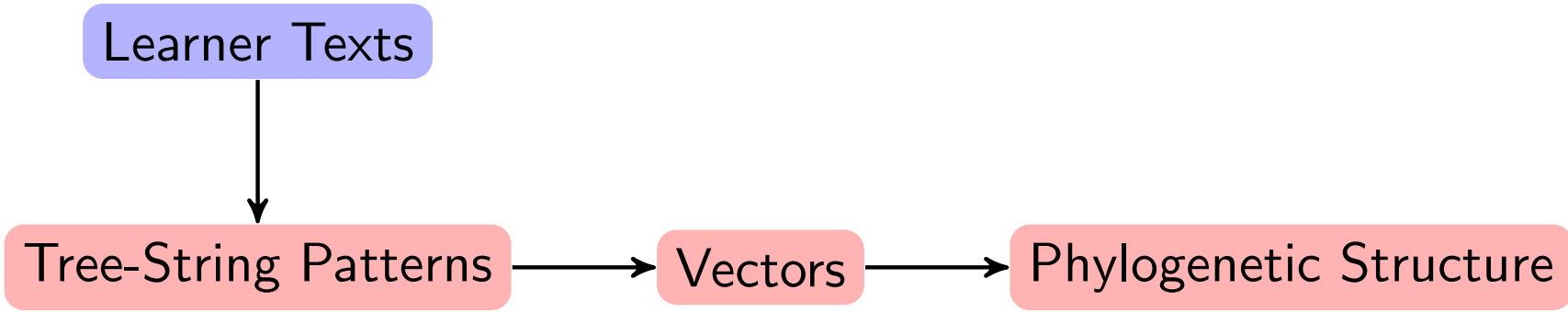


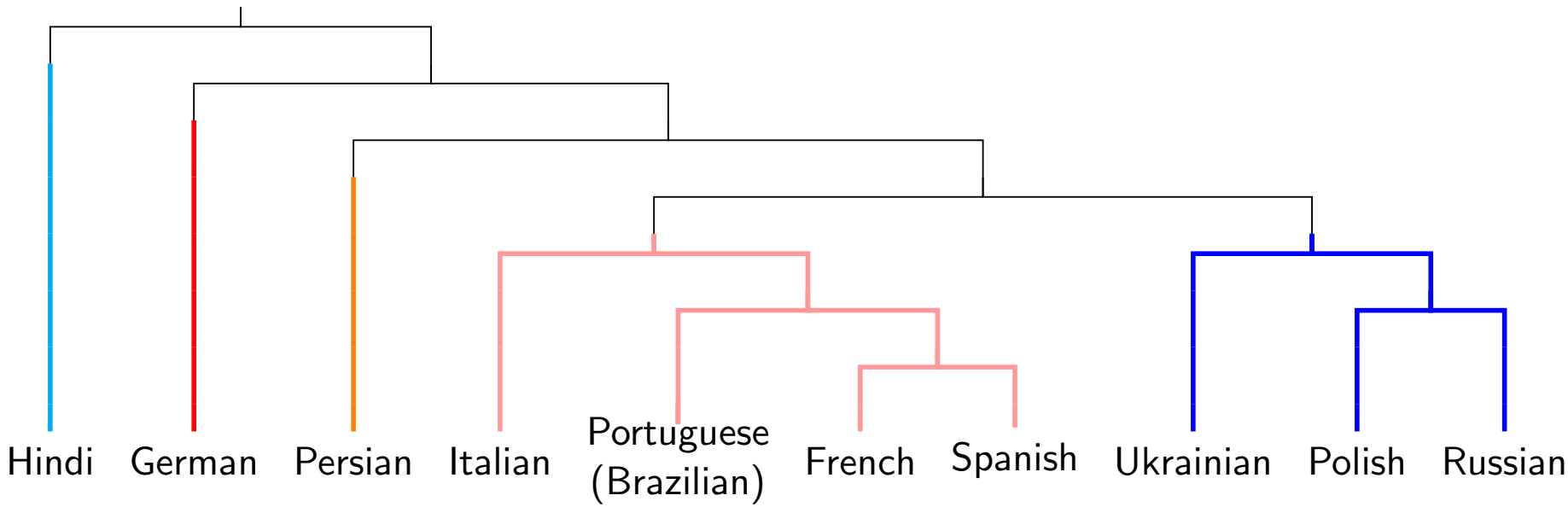
VP(x0:VB x1:ADVP x2:NP) → x1 x0 x2





Using patterns

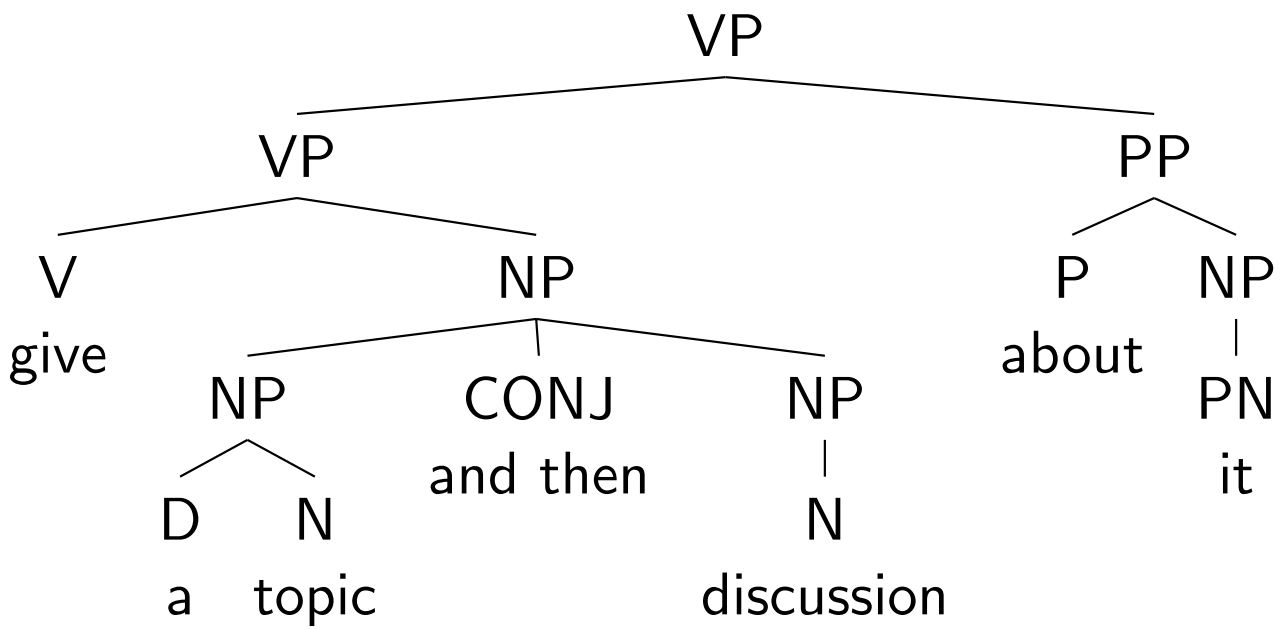




Zhao et al. (2020); arXiv:2007.09076

Interface Hypothesis

Language structures involving an interface between different language modules, like **syntax–semantics interface** and **semantics–pragmatics interface**, are less likely to be acquired completely than structures that do not involve this interface; see e.g. Sorace 2011



Literal meaning versus intended meaning

Literal Meaning

conventional meaning
sentence meaning

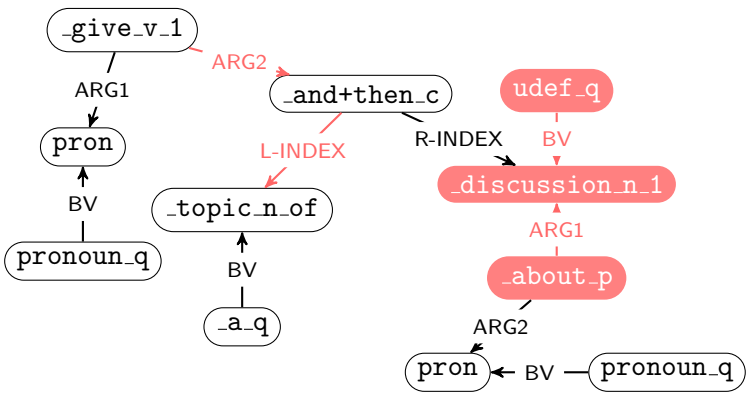
linguistic code features

often
≠

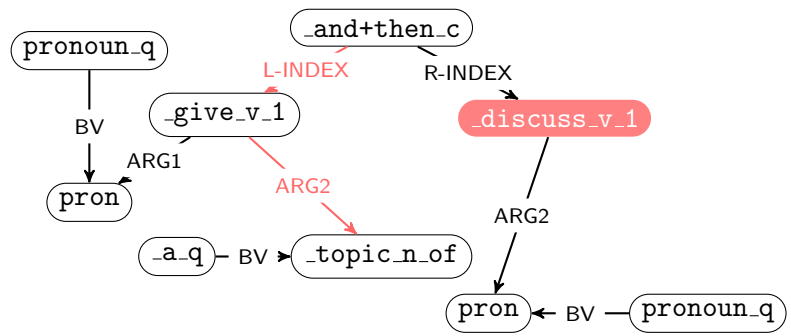
Intended Meaning

speaker meaning
interpretation

author's intention



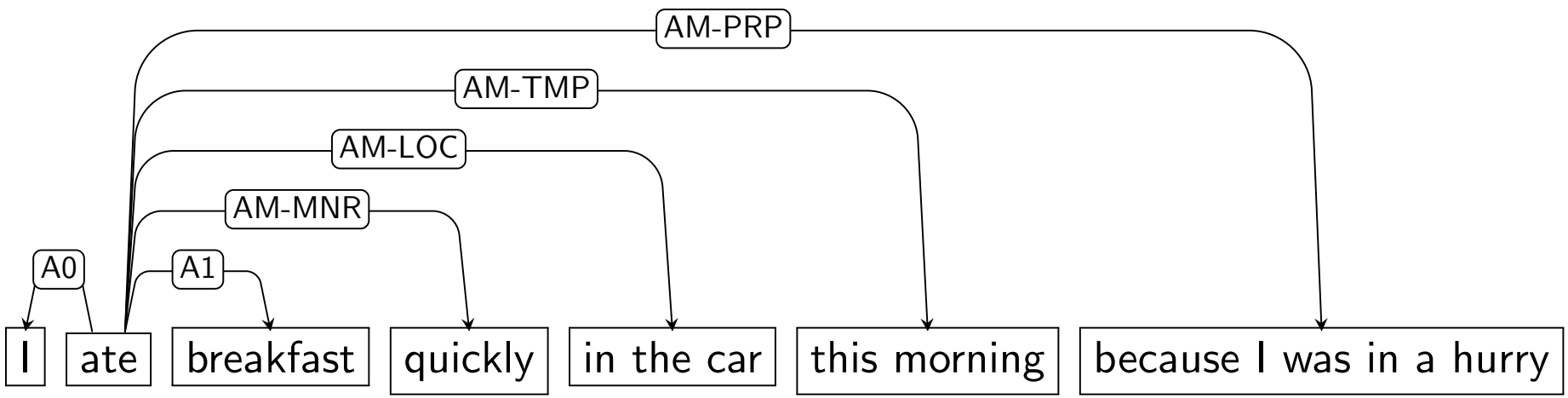
give a topic and then discussion about it.



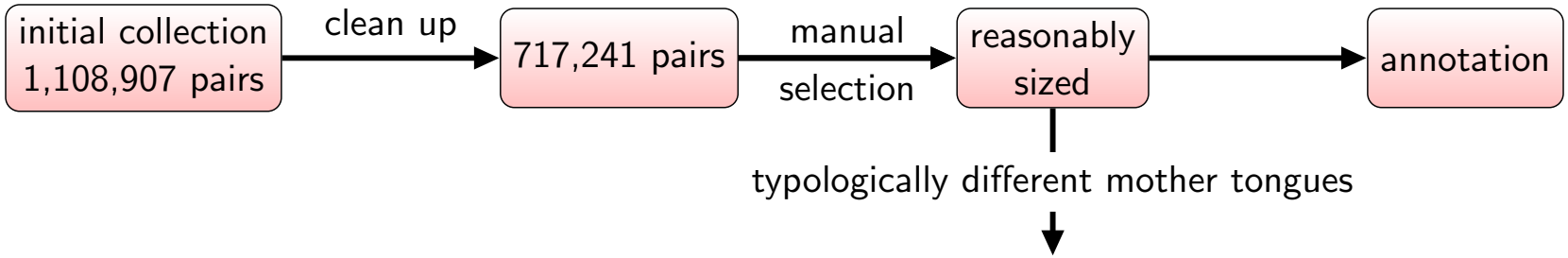
Give a topic and then discuss it.

Semantic Role Labeling

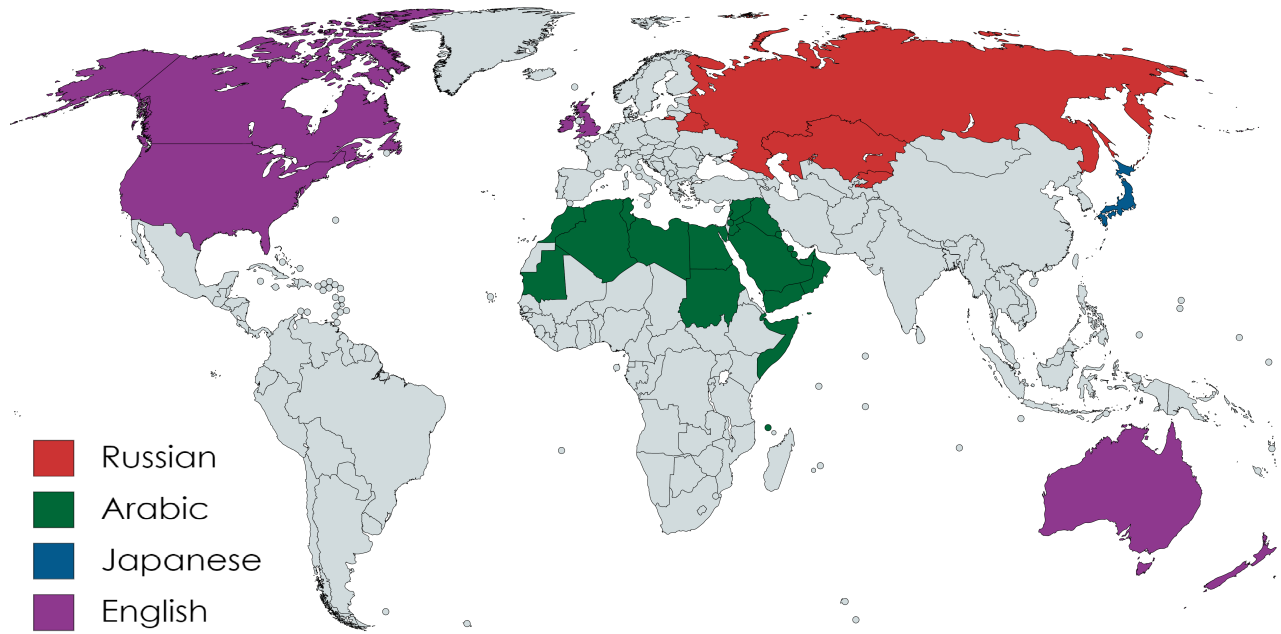
- **Argument (AN):** Who did what to whom?
- **Adjunct (AM):** When, where, why and how?



Data source

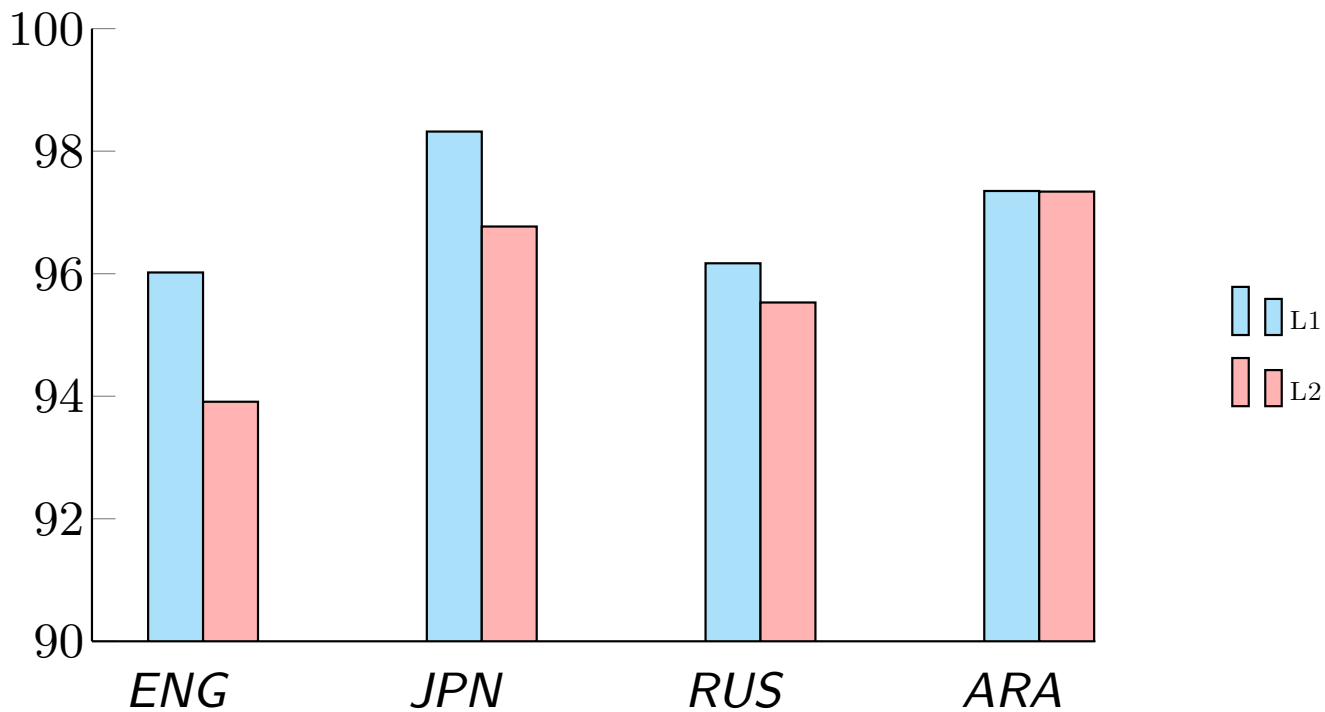


Chinese	Sino-Tibetan
Russian	Slavic
Arabic	Semitic
Japanese	?
English	Germanic

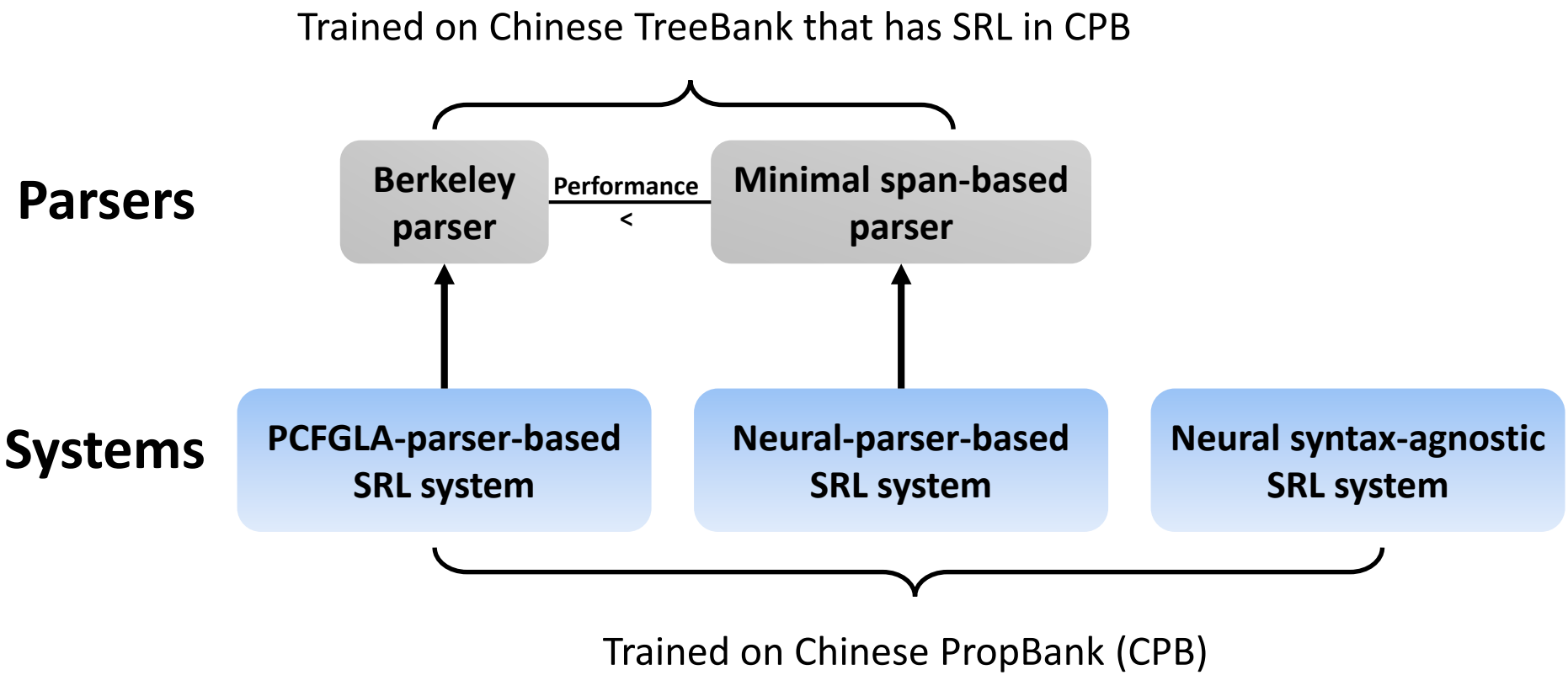


Inter-annotator agreement

- **Annotator:** two students majoring in linguistics
- **The first 50-sentence trial set:** adapting and refining the Chinese PropBank secification
- **The rest 100-sentence set:** reporting the inter-annotator agreement



Three SRL systems



Syntax-agnostic SRL

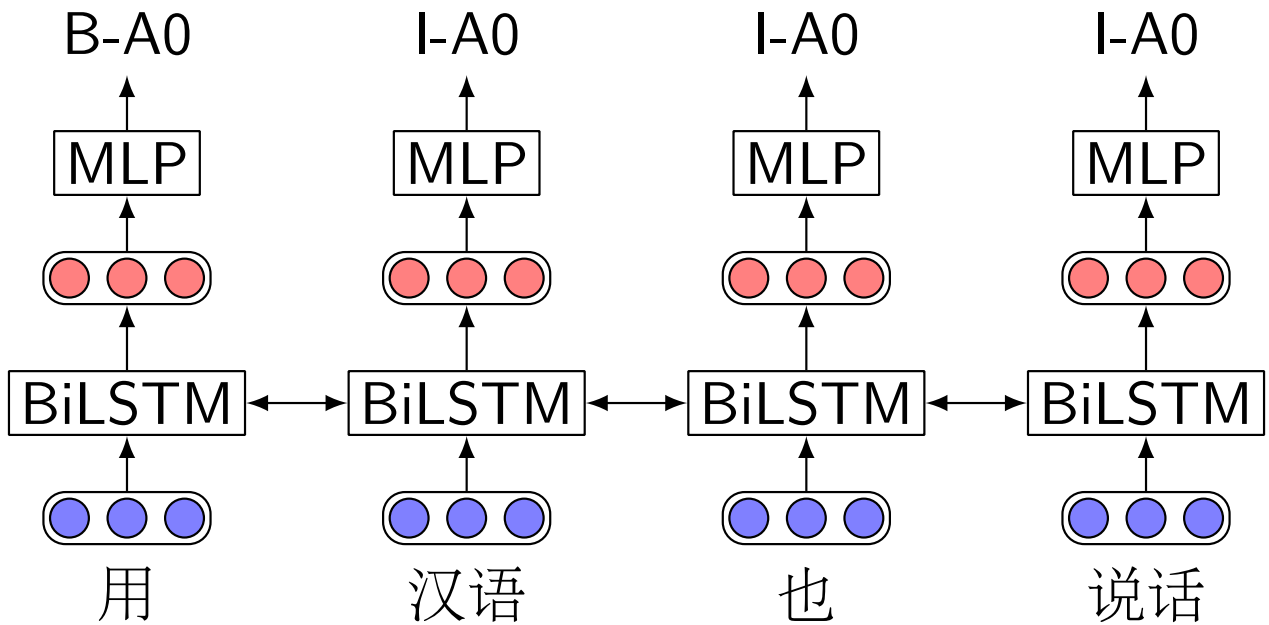
B-A0	I-A0	I-A0	I-A0	I-A0	B-AM	I-AM	I-AM	I-AM	B-AM	REL
用	汉语	也	说话	快	对	我	来	说	很	难

Syntax-agnostic SRL

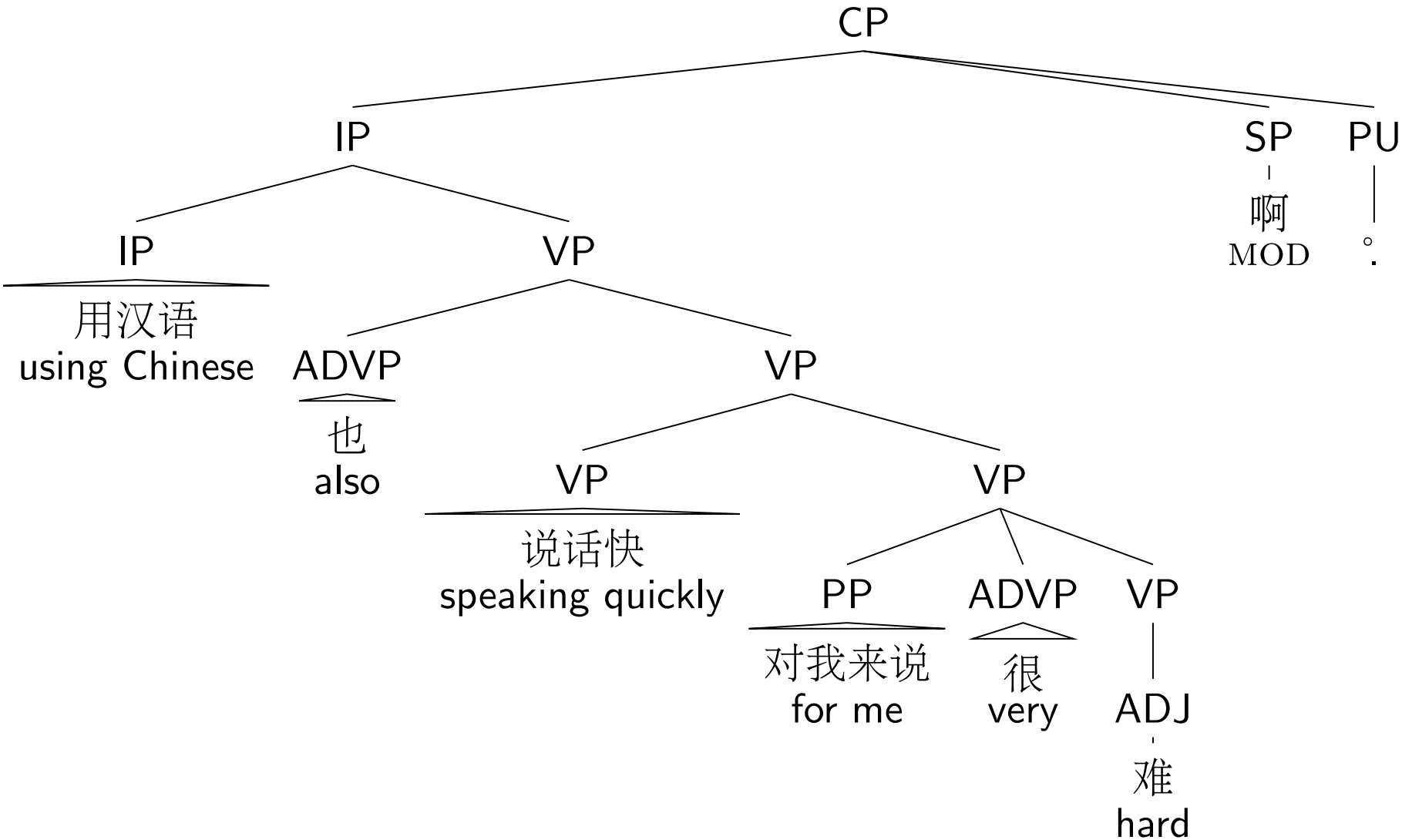
B-A0	I-A0	I-A0	I-A0	I-A0	B-AM	I-AM	I-AM	I-AM	B-AM	REL
用	汉语	也	说话	快	对	我	来	说	很	难

Syntax-agnostic SRL

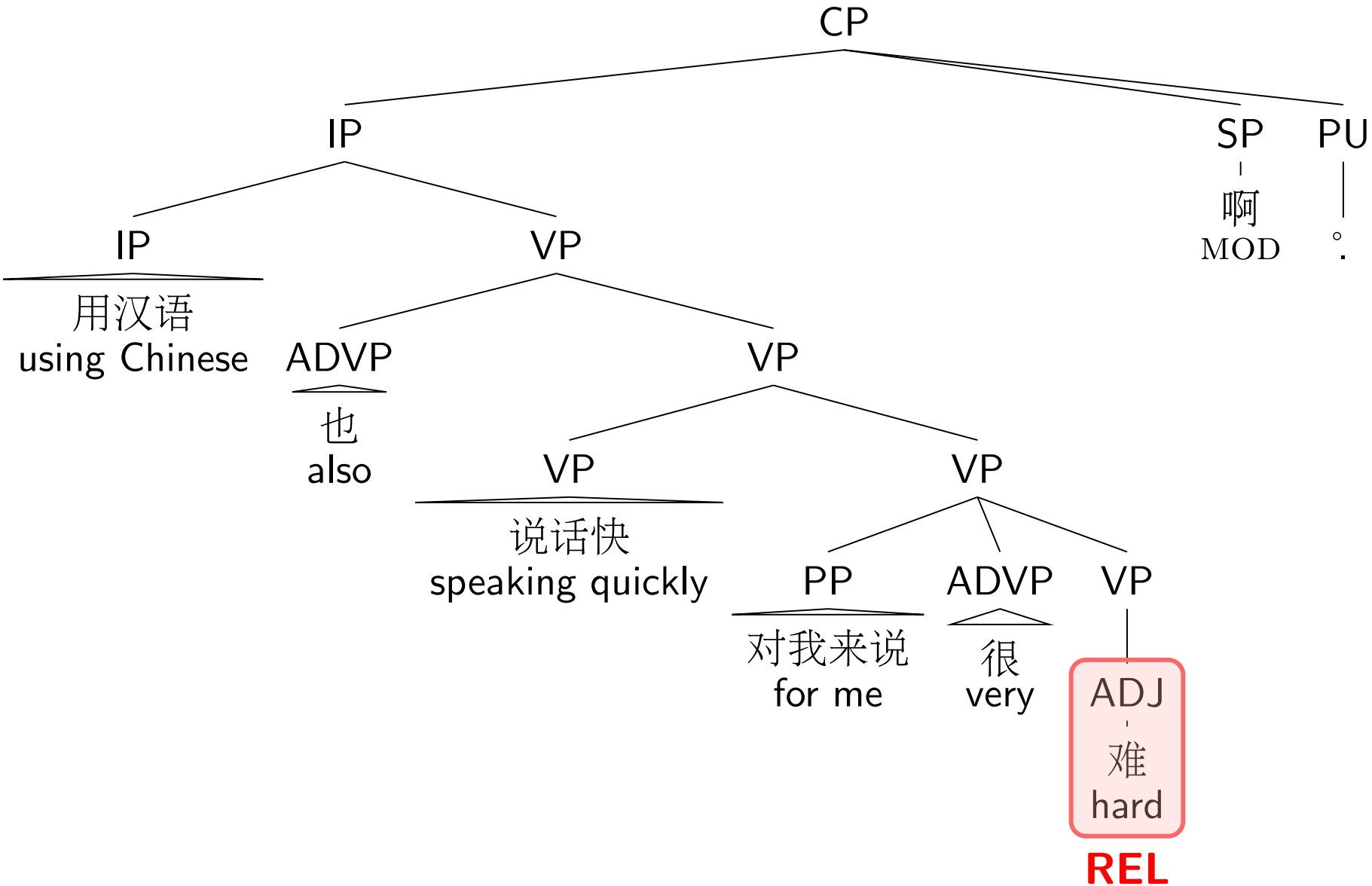
B-A0	I-A0	I-A0	I-A0	I-A0	B-AM	I-AM	I-AM	I-AM	B-AM	REL
用	汉语	也	说话	快	对	我	来	说	很	难



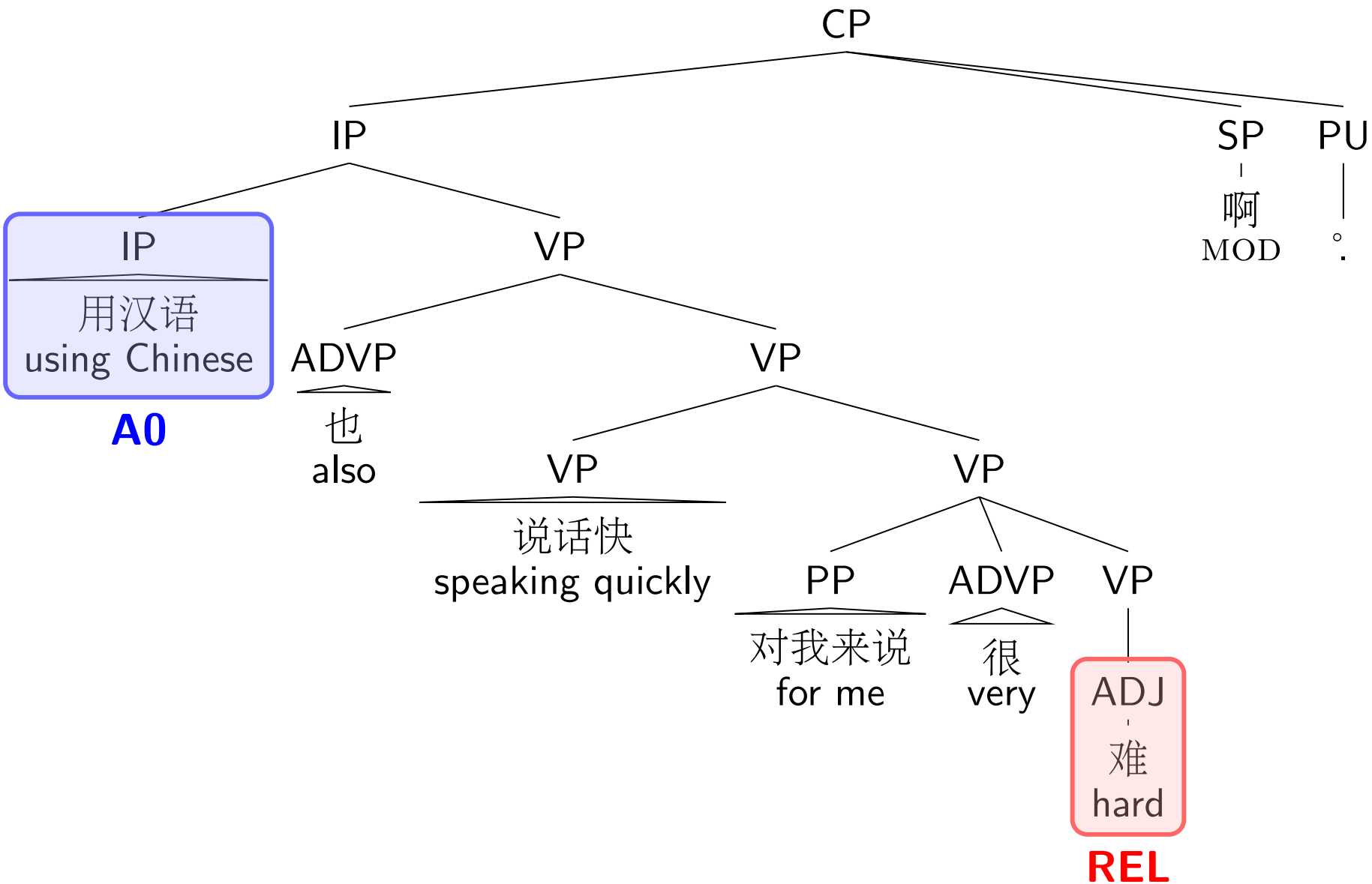
Syntax-based SRL



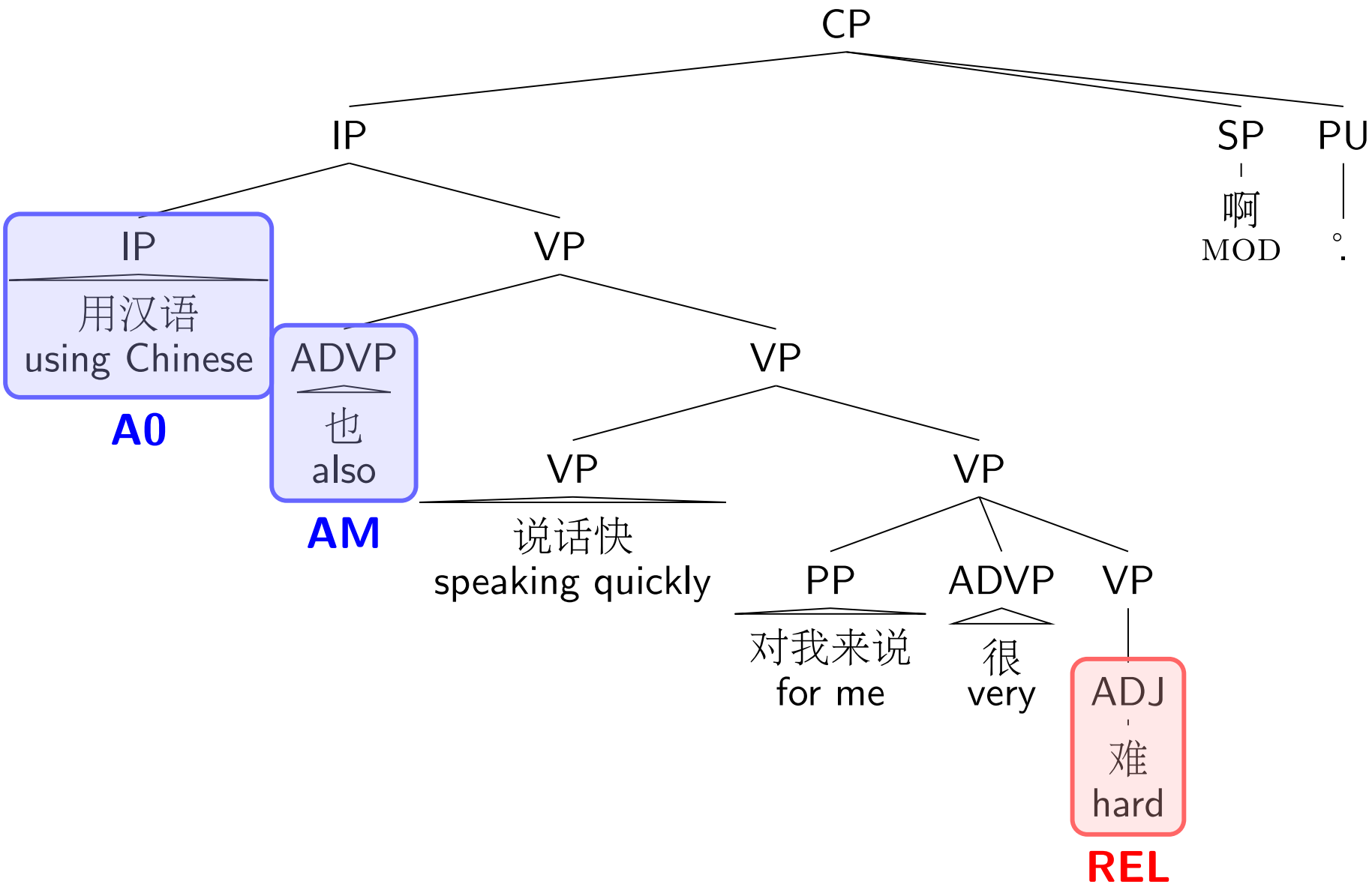
Syntax-based SRL



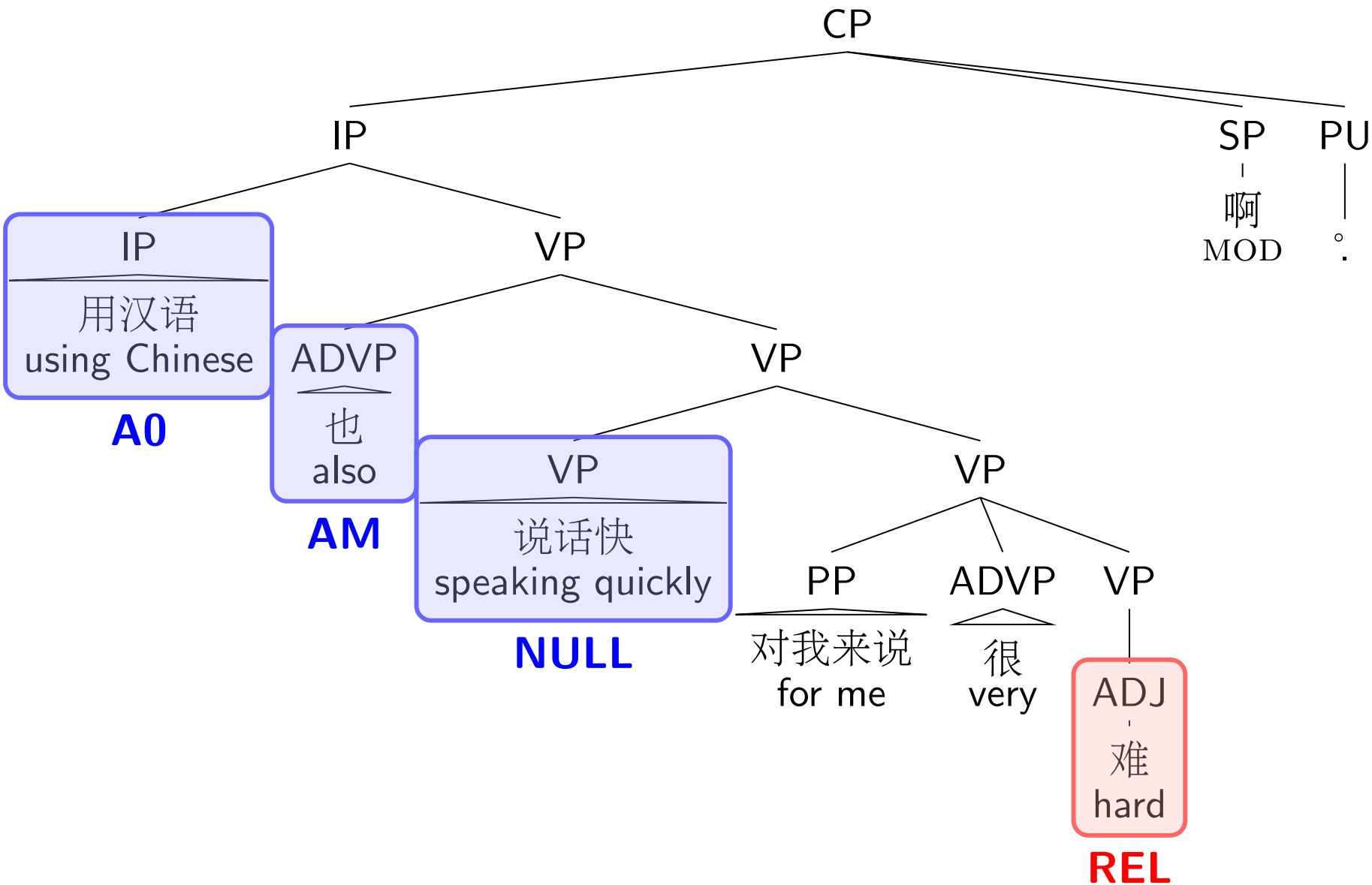
Syntax-based SRL



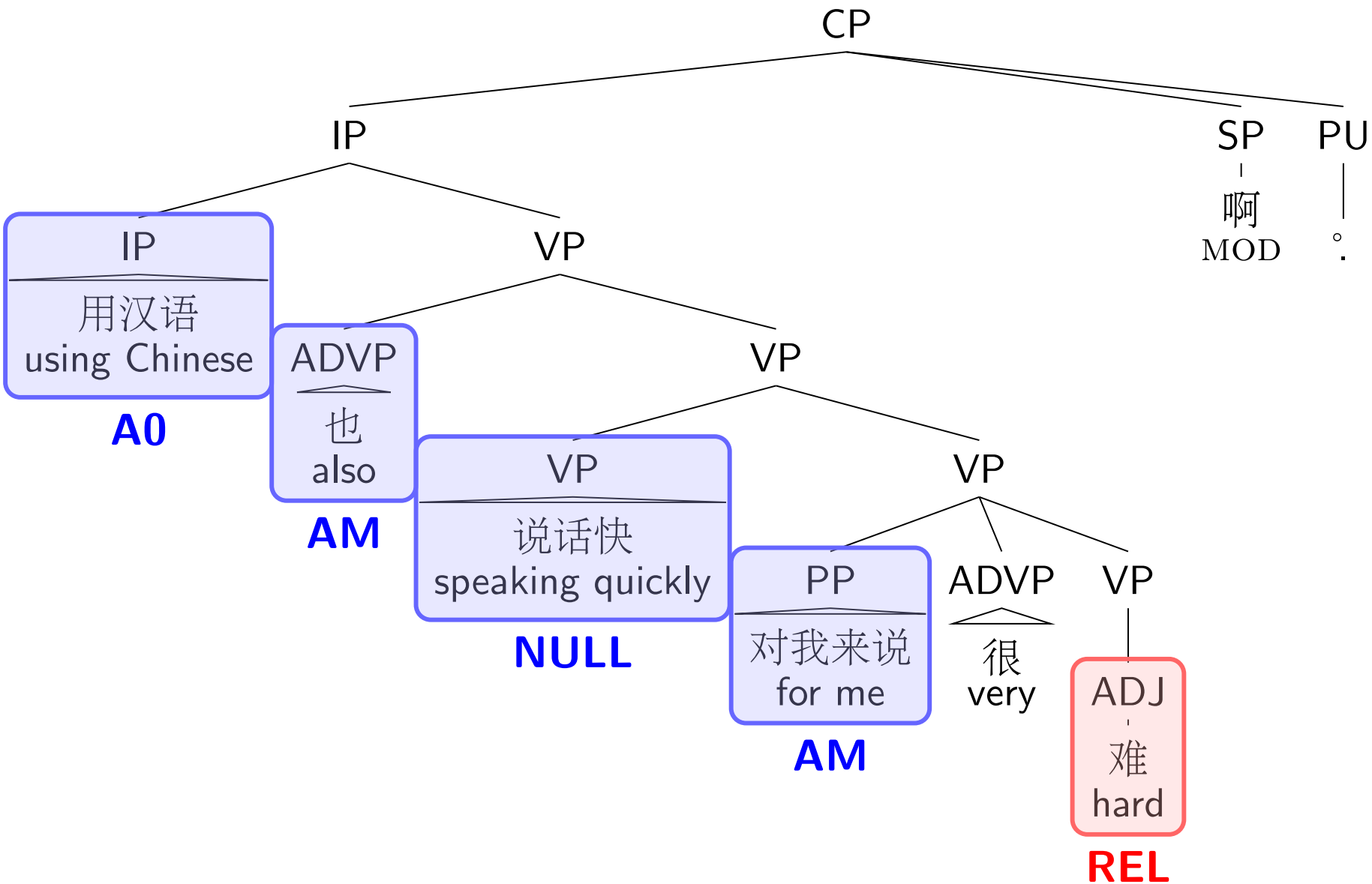
Syntax-based SRL



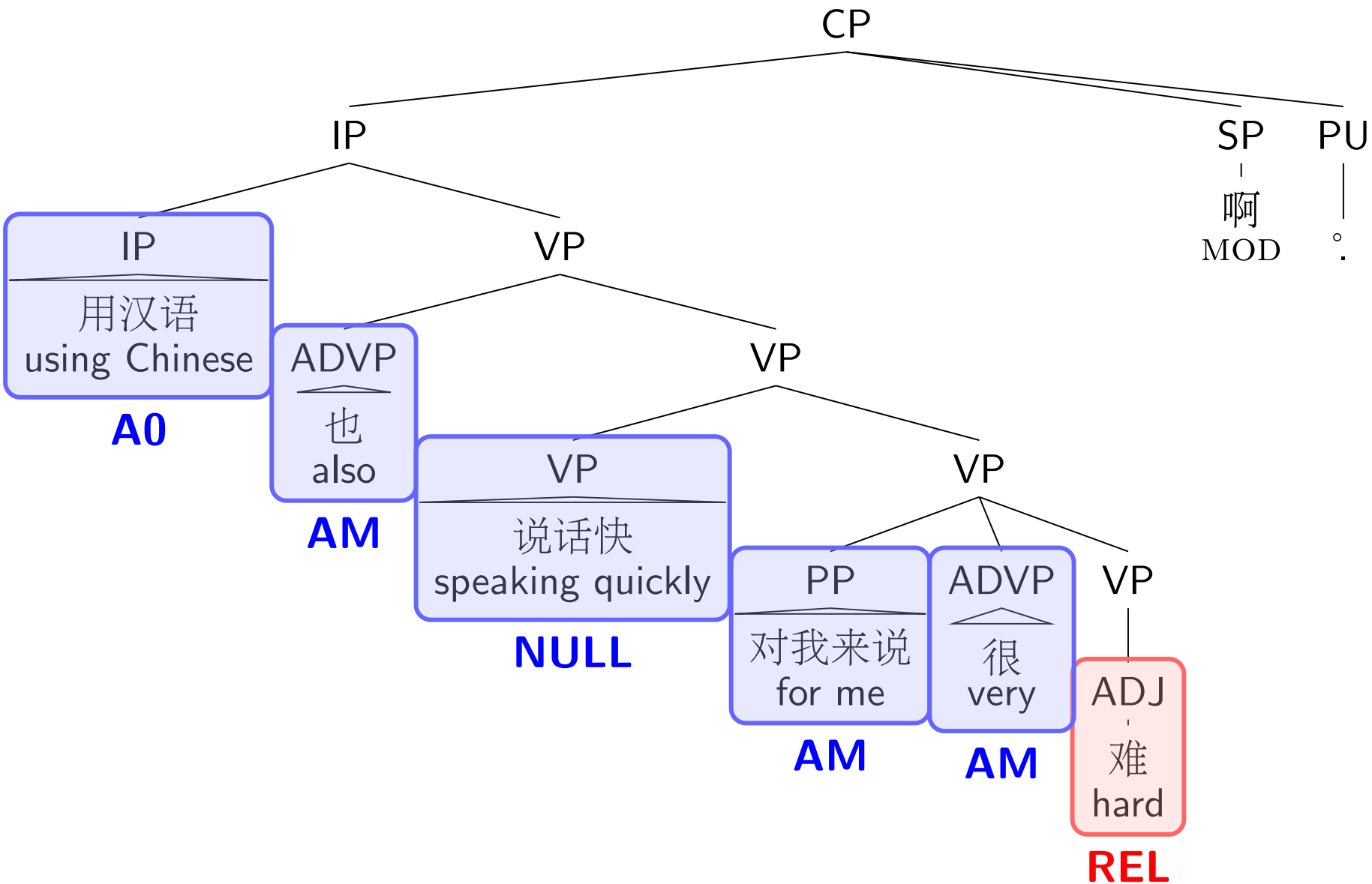
Syntax-based SRL



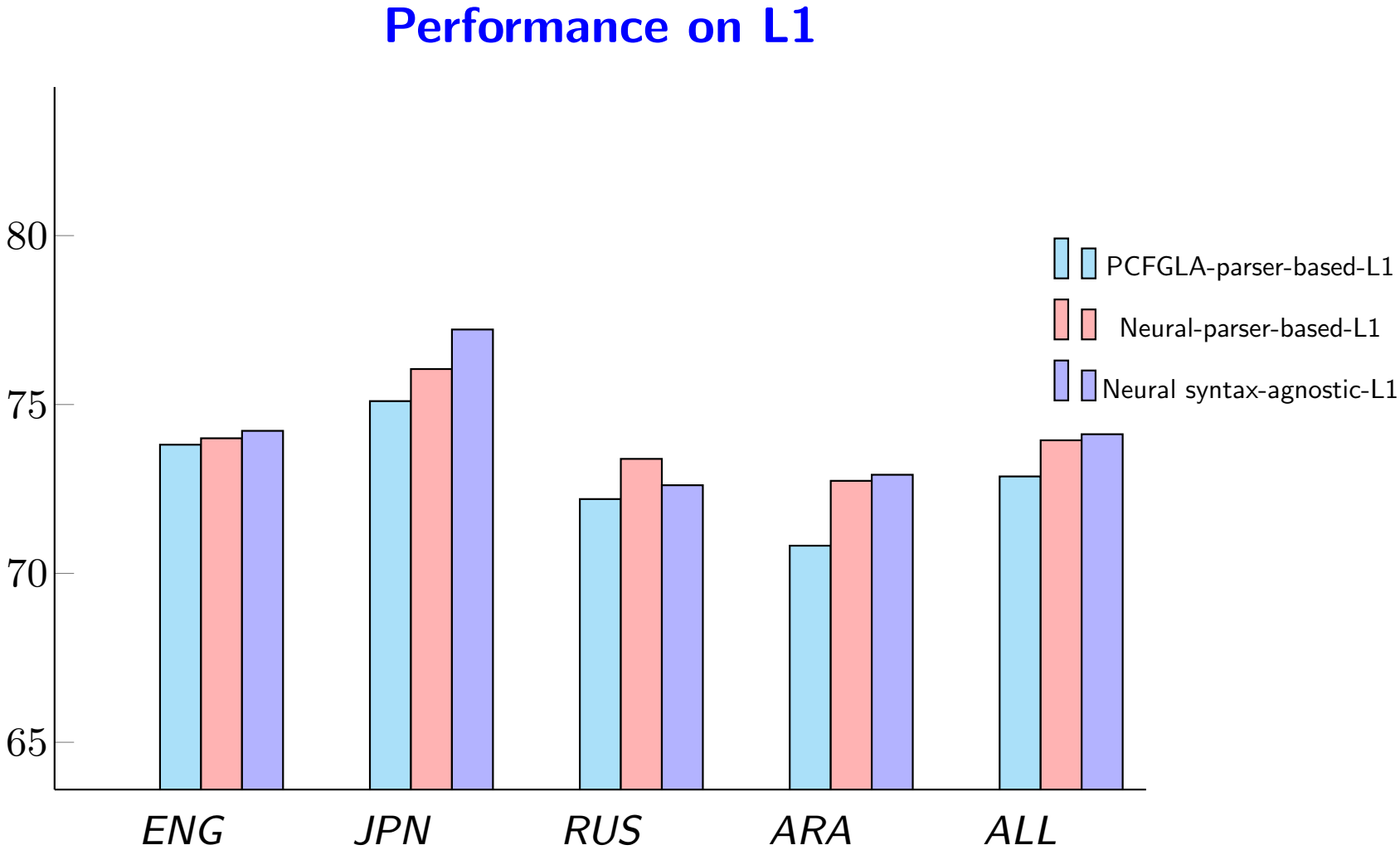
Syntax-based SRL



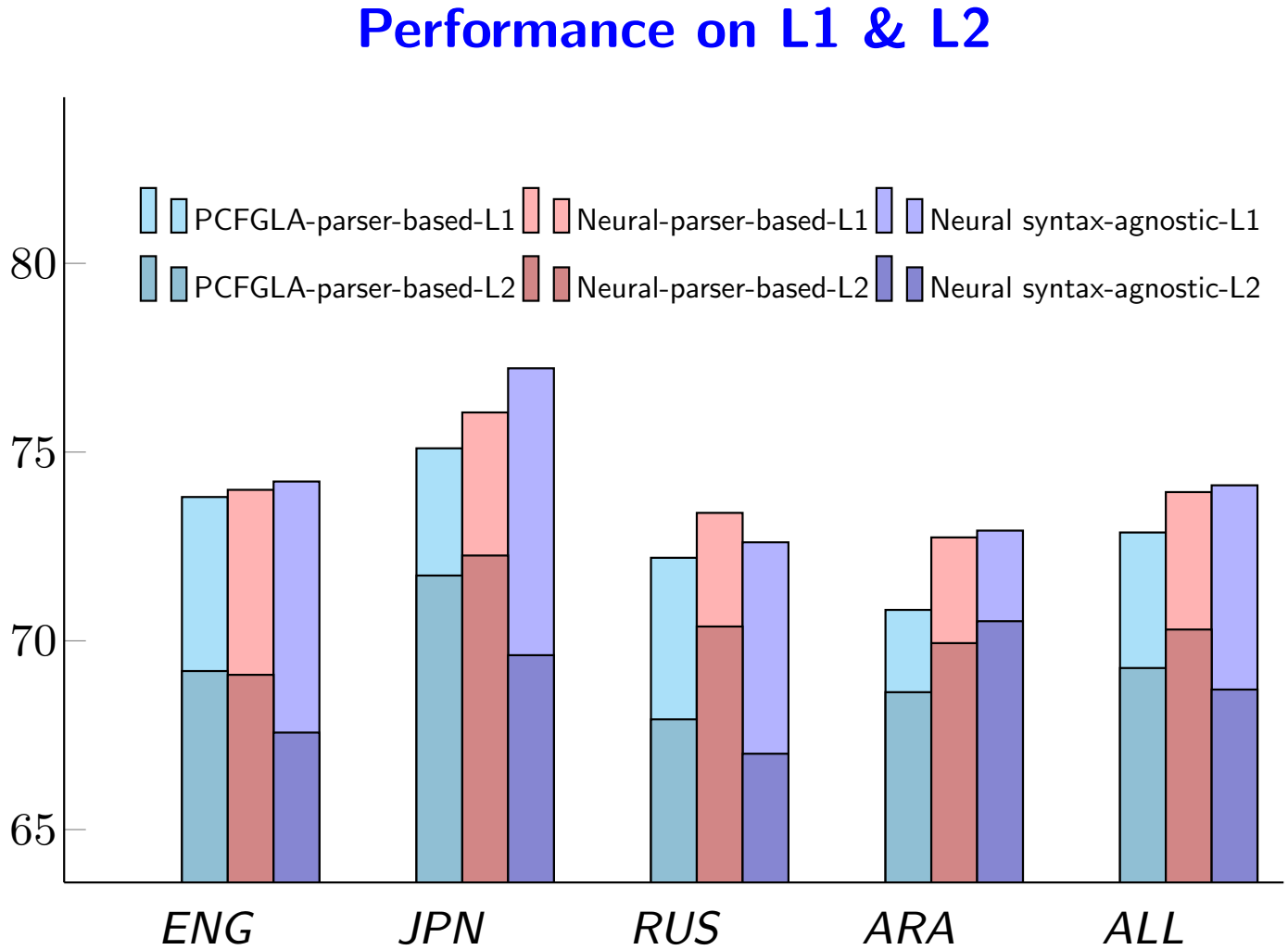
Syntax-based SRL



Evaluation and findings

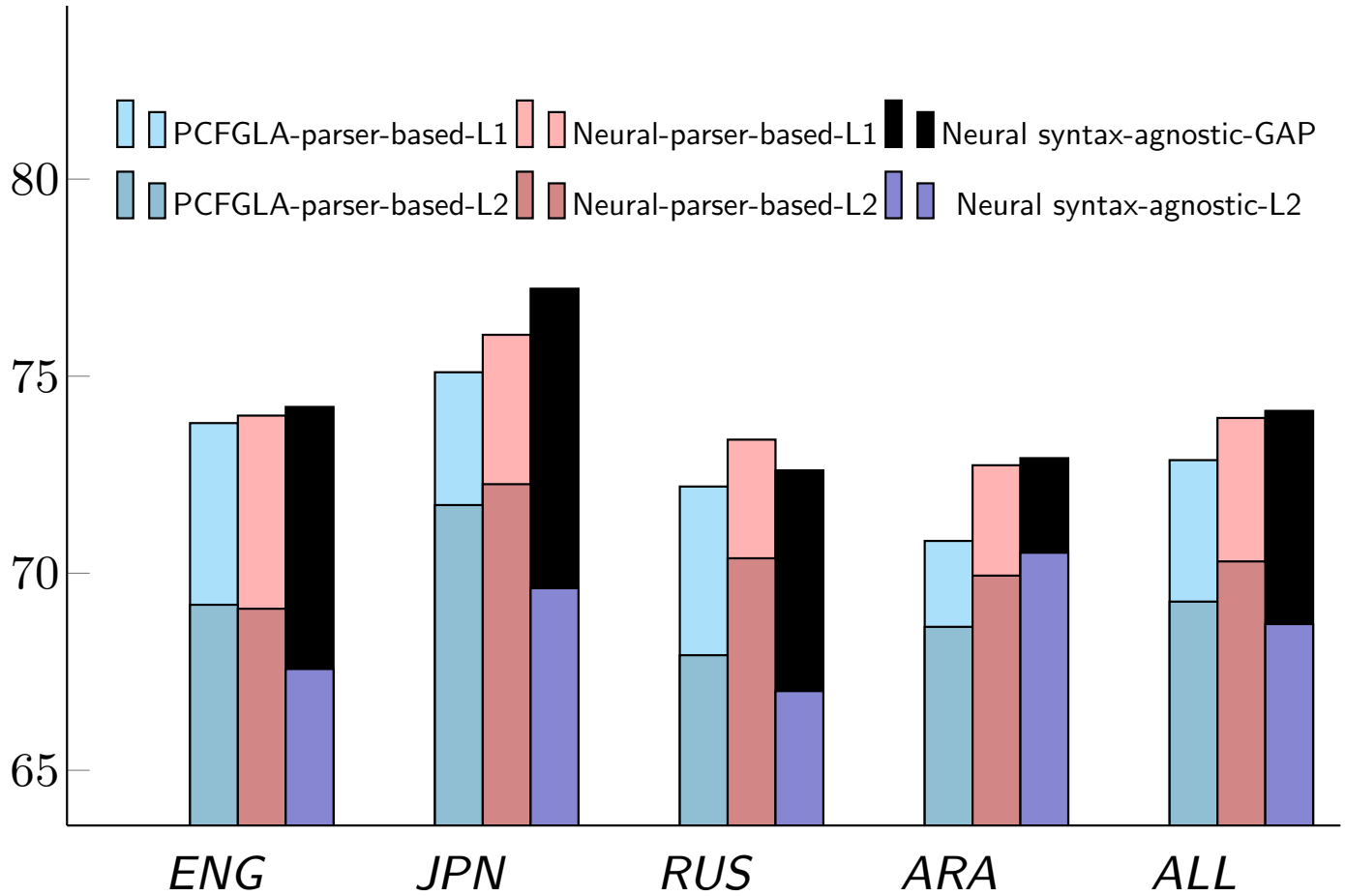


Evaluation and findings



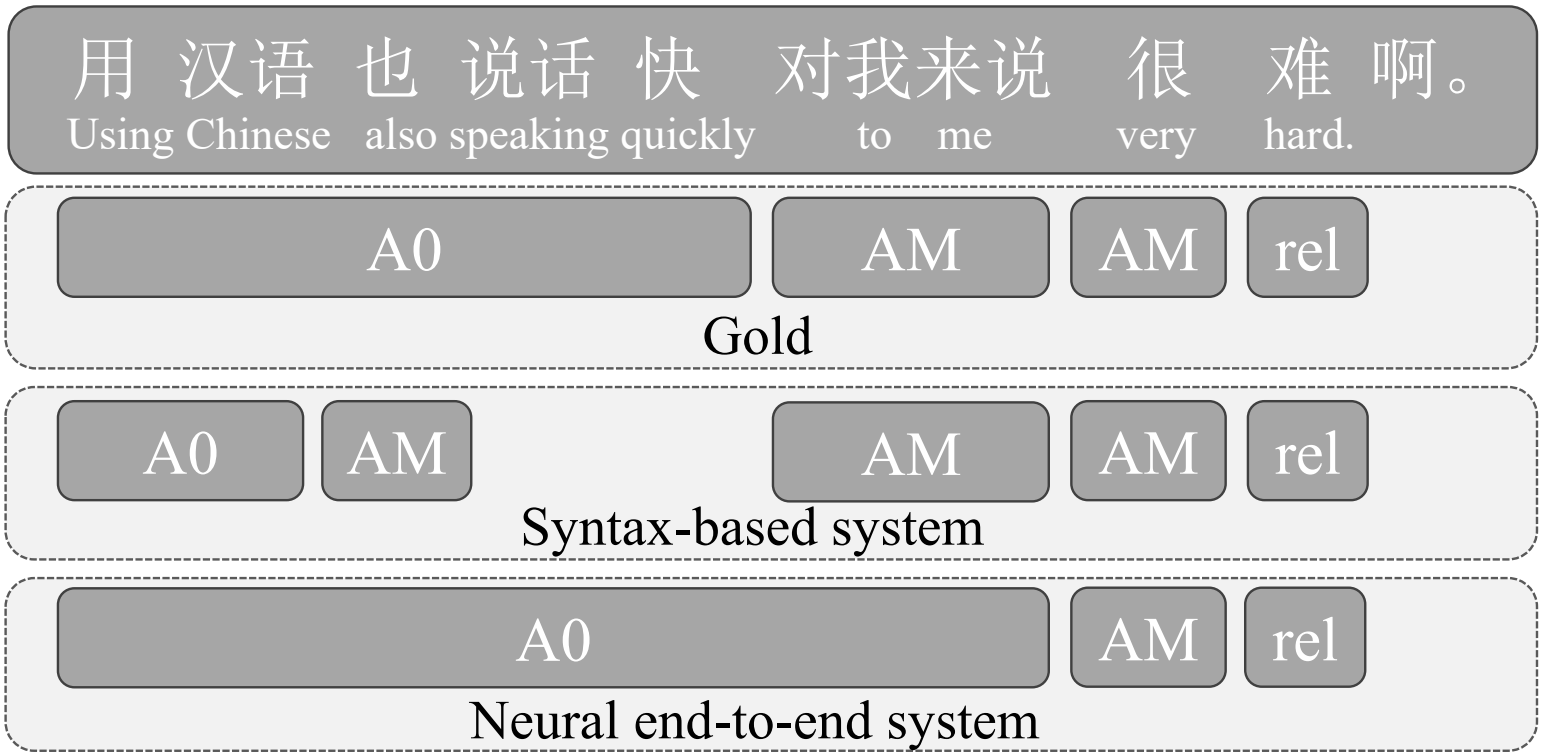
Evaluation and findings

Performance on L1 & L2



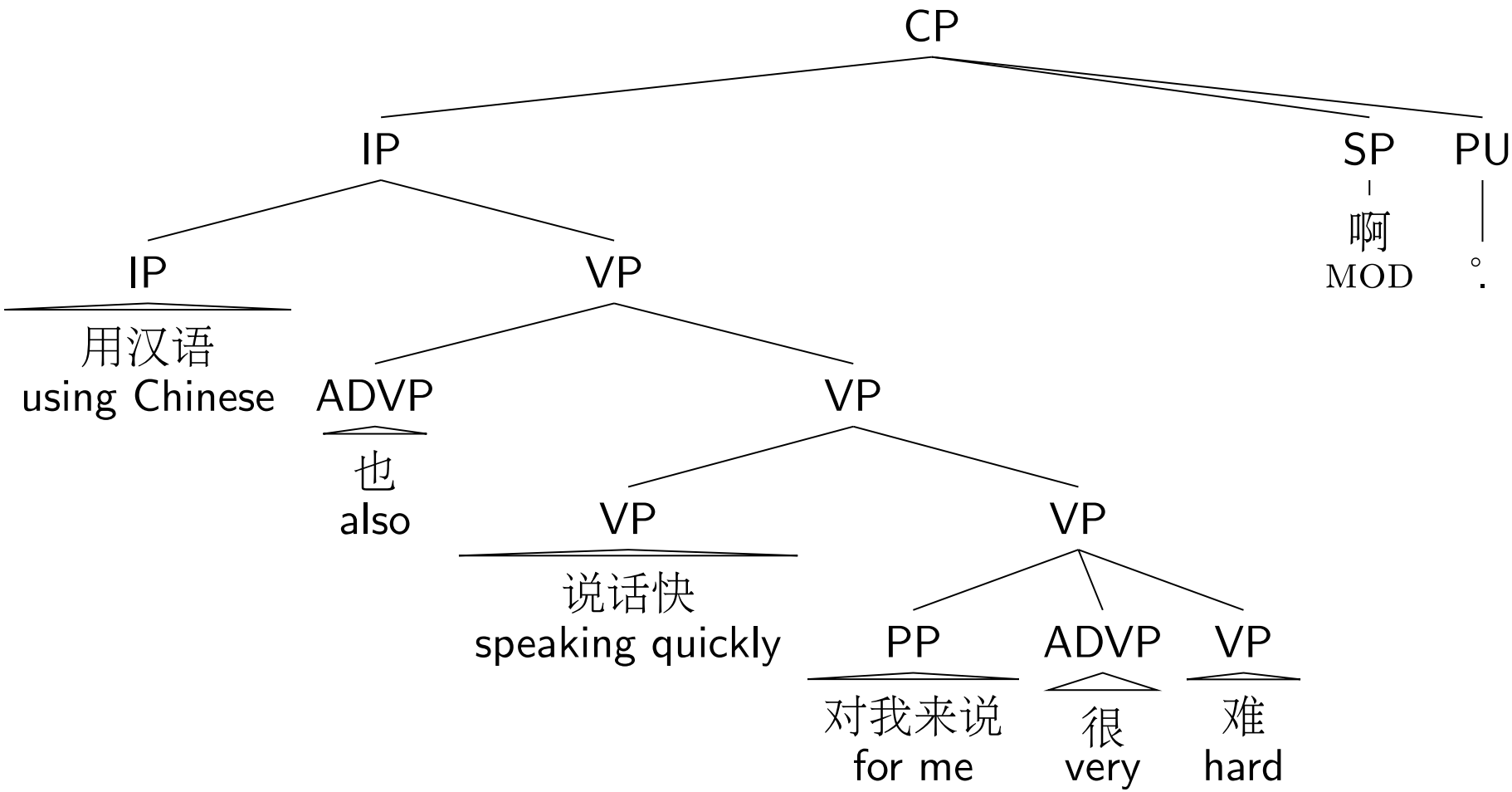
The syntax-based systems are more robust when handling learner texts.

Why syntactic analysis is important?

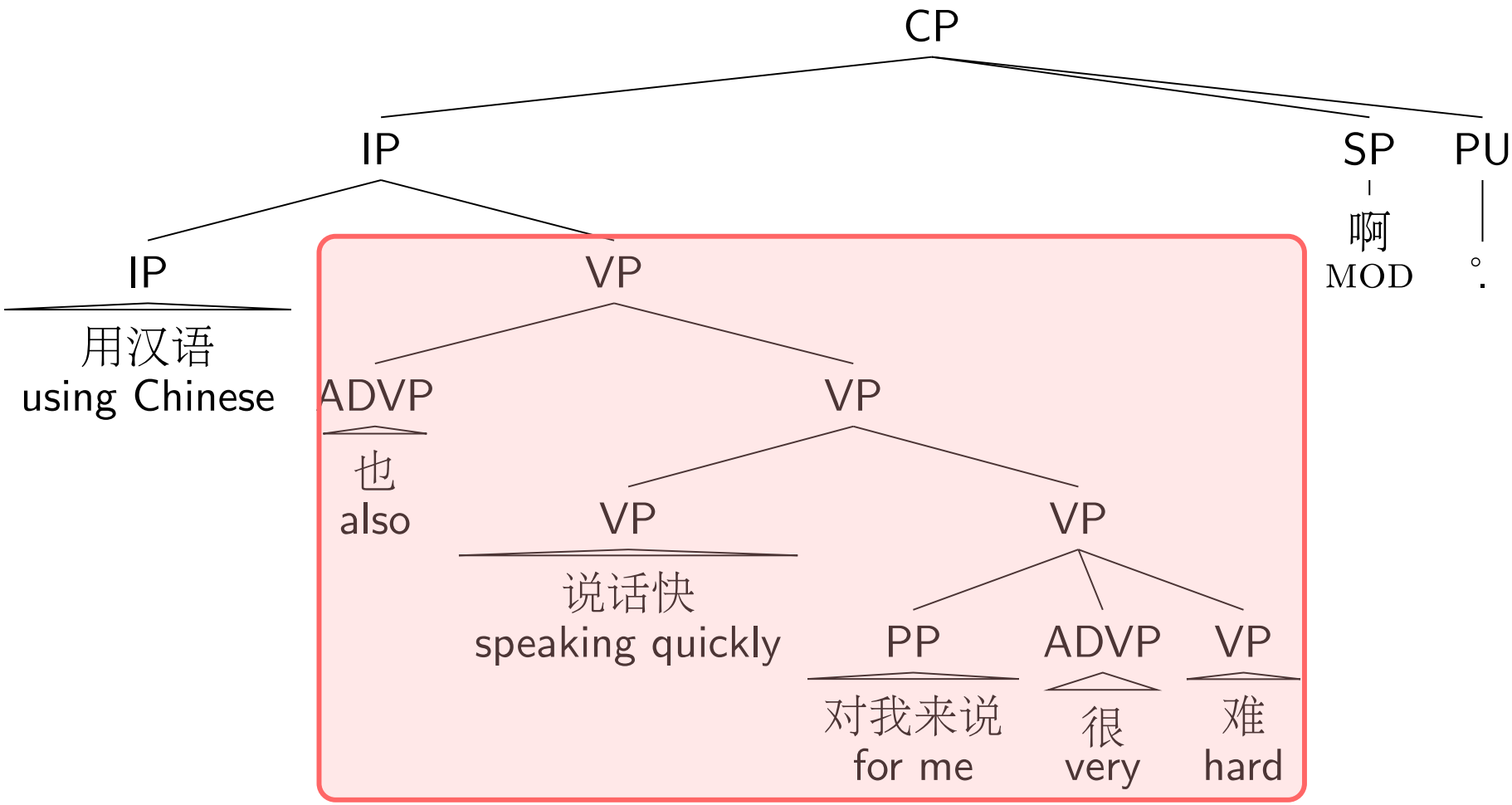


It is very hard for me to speak Chinese quickly.

Why syntactic analysis is important?

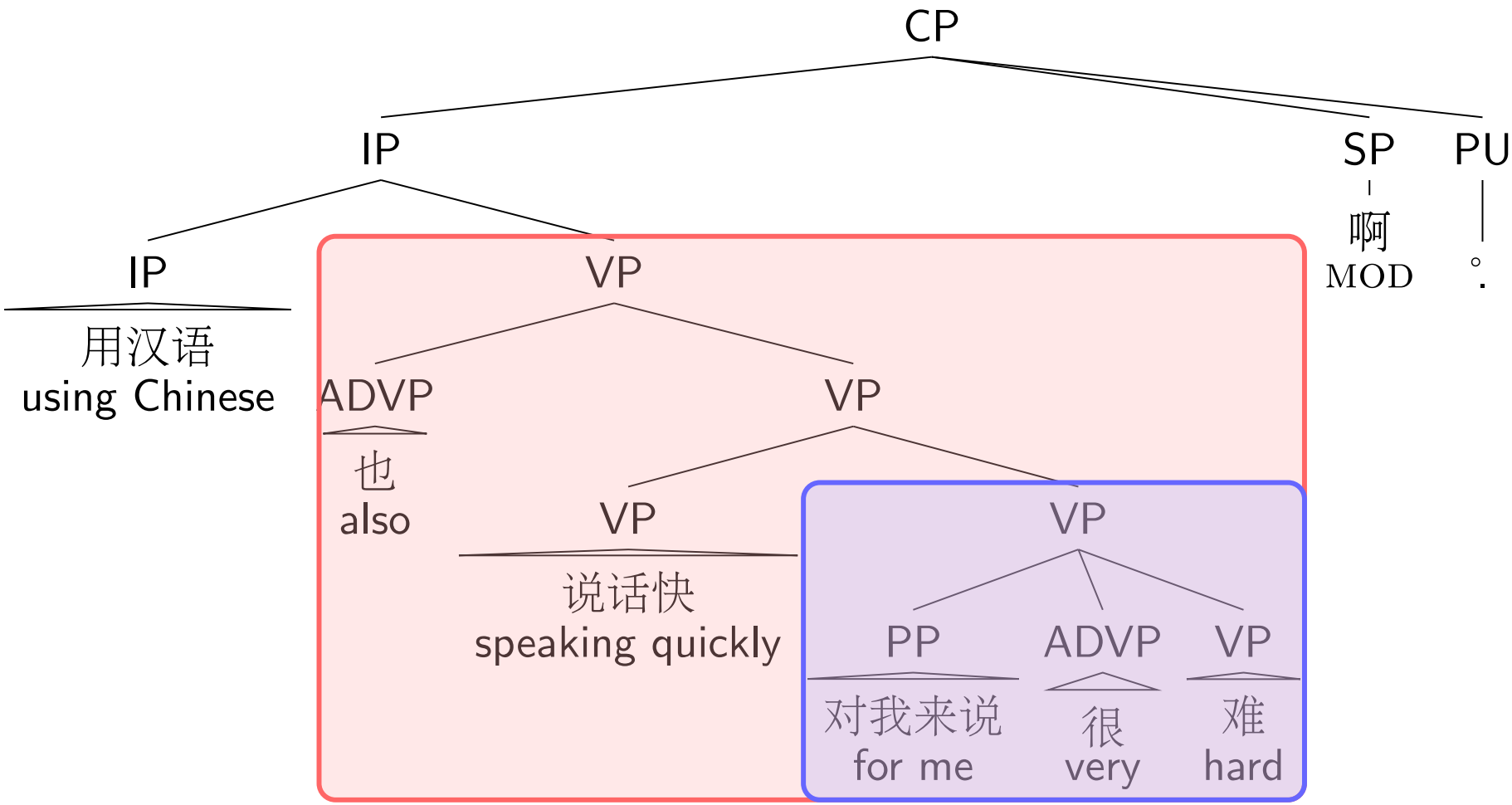


Why syntactic analysis is important?



Though the whole structure is *bad*,

Why syntactic analysis is important?



Though the whole structure is *bad*, some parts may be *good*.