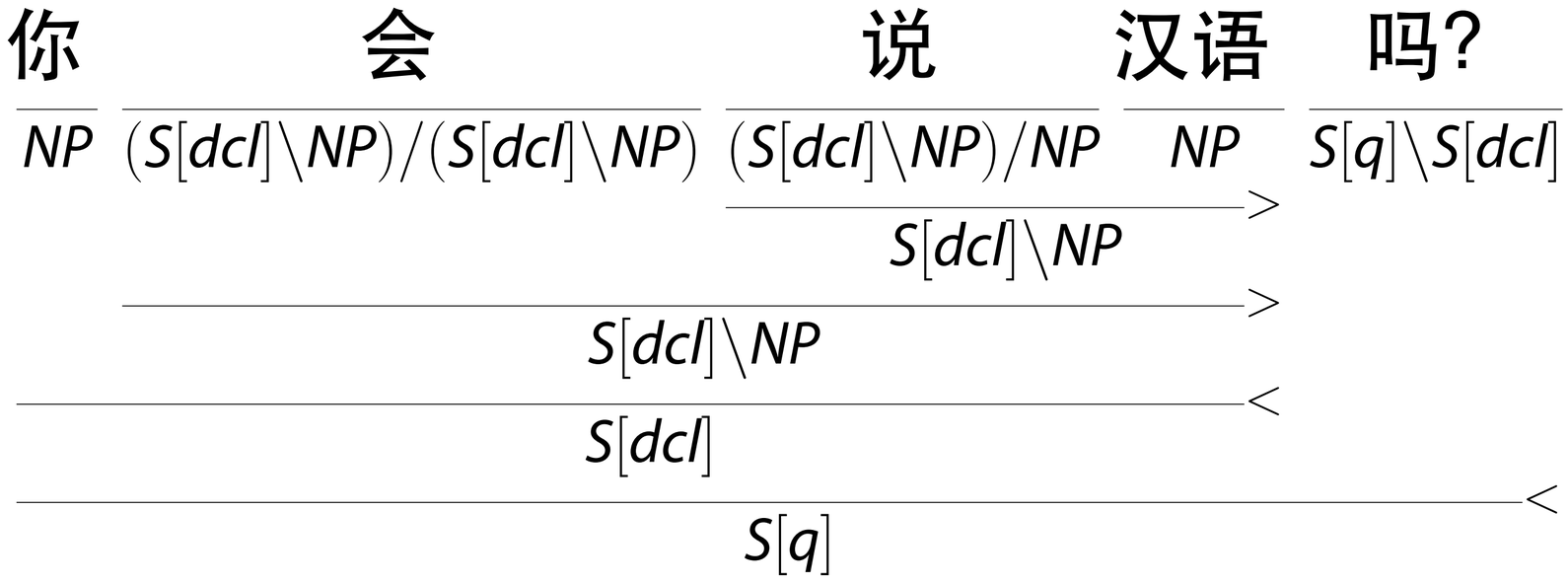
CHINESE CCGbank

Building fast, accurate, rich parsers for the Chinese language

Daniel Tse
School of IT, FACULTY OF ENGINEERING & INFORMATION TECHNOLOGIES



WHAT WE'VE DONE AND WHY

- **China has more Internet users than the entire population of the United States.**
- As the quantity of Chinese text grows, so does the need for efficient, automatic text processing tools.
- **Building those tools demands data. *Lots of data.***
- A theory of syntax and semantics called **Combinatory Categorial Grammar (CCG)** can be used to build a rich description of Chinese grammar.
- We have a way of getting ***lots of data*** for CCG in the Chinese language.
- **Introducing *Chinese CCGbank:* 0.75 million Chinese words, annotated for CCG, ready for deep, efficient, automatic Chinese text processing.**

CHINESE LANGUAGE SYSTEMS NEED CHINESE DATA

- Modern natural language processing systems often rely on the availability of **large, wide-coverage collections of data** to learn a model of natural language.
- There are many theories of **how to encode the grammar of a language, like Chinese.**
- We have shown that a theory called **Combinatory Categorial Grammar (CCG)** accounts for Chinese syntax & semantics concisely and effectively.

THE MILLION DOLLAR QUESTION

Approach 1:

- Spend millions of dollars and 100,000 man hours **developing your own dataset** for a new language, employing linguists and computational linguists.

Approach 2:

- Don't reinvent the wheel – **take an existing dataset, and transform it** into a new resource.

CCG LEADS TO FAST, DEEP, RICH PARSING

- The goal of a parser is to uncover the **hidden structure of language**.
- The structure can be used to **build search engines that understand documents**.
- CCG has been used successfully as the core of a parser responsible for **state-of-the-art parsing in English**.
- We want to bring the power and flexibility of that same parser to Chinese.

THE APPROACH

1. Take **PENN CHINESE TREEBANK**, an existing database of Chinese text annotated for syntax
2. **Design an analysis** of Chinese syntax through CCG
3. **Automatically convert** Penn Chinese Treebank derivations into CCG derivations
4. Create and train **rich, efficient Chinese parsers based on CCG** using Chinese CCGbank

THE RESULTS

1. We have the **first analyses of Chinese syntax using the power of CCG**.
2. We have an automatic system for converting Penn Chinese Treebank derivations into CCG derivations.
3. We have **Chinese CCGbank, a database of 0.75 million words of Chinese text** annotated with their syntax & semantics.

CHINESE CCGbank: ANALYSIS

HOW MUCH OF THE ORIGINAL CORPUS HAVE WE PRESERVED?

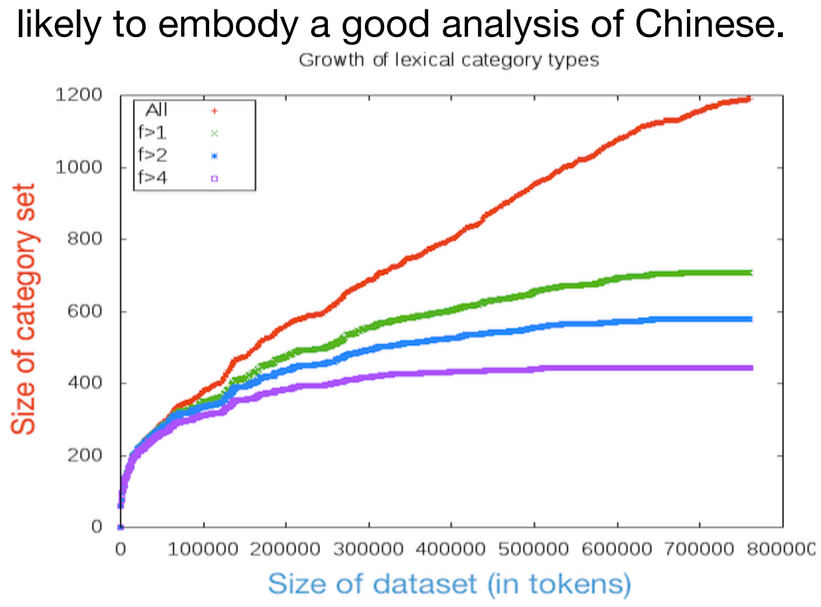
Original corpus	28295 derivations
Our algorithm produces	28227 (99.76%)
Filtering out infrequent categories	27759 (98.11%)
Filtering out infrequent rules	26680 (94.30%)

IS THE ORIGINAL CORPUS BIG ENOUGH TO YIELD A GOOD MODEL OF CHINESE?

We graph the size of the **CCG tag set** against the **size of the corpus (in words)**.

If the resulting graph shows an **upward trend** then the original corpus is likely too small.

Ignoring words than only occur once, the trend flattens out, so Chinese CCGbank is likely to embody a good analysis of Chinese.



WHAT'S NEXT?

- We sidestepped the cost of developing your own dataset by transforming an existing one.
- We can harness CCG, a powerful, efficiently parseable theory of grammar, to analyse Chinese.
- We have **Chinese CCGbank**, a large database of Chinese text annotated with syntax & semantics.
- We are ready to build **efficient, deep parsers using Chinese CCGbank**, to meet the vast text processing needs which Chinese will demand in the years to come.