# Implementing Numerical Expressions in Mandarin Chinese

Dan Flickinger & Chunlei Yang

(Stanford University & Shanghai International Studies University)

# Outline

1. Introduction

2. Syntax of numerical expressions

      2.1 Regular numerical expressions

      2.2 零 (zero) lexical types

      2.3 "Irregular" numerical expressions

3. Semantics of numerical expressions

4. Proficiency check with test suits

5. Conclusion

# 1. Introduction

- Started with Zhong [|] Grammar

- Major changes in zhong-lextypes.tdl and zhong-letypes.tdl.

- Some changes in zhong.tdl, e.g.,
  - Modified "frag-np-rule" to raise a numerical expression to a frag-np position;
  - Added "num-nominal-phrase" to enable a numerical expression to function like a nominal phrase;

# 1. Introduction

- Enriched lexicon with more handcrafted entries, e.g., numerical units and various entries based on test suites;

- The basic numerals from 一 (one) to 九 (nine) in lex-numbers.tdl are kept in use.

# 1. Introduction

- Identification of the phenomenon to focus on:
  - Build and batch parse a test suite;
  - The results show that 25% of the parse failures include some numerical expression;
  - The grammar does not parse a numerical expression which is syntactically and semantically composite, e.g., 二十 (two ten; "20").

# 1. Introduction: basic facts

- Three components of Chinese numerical expressions
  - Basic numerals: from 1 through 9
  - Numerical units: e.g.,十 (ten), 百 (hundred), 千 (thousand), 万 (ten-thousand), 亿 (hundred-million) …
  - 零 (zero)
- An example:

四　　千　　三　　　百　二 十 一

four thousand three hundred two ten one

"4,321 / four thousand three hundred twenty one"

# 1. Introduction: contrastive units

| digit scale: | $10^1$ | $10^2$ | $10^3$ | $10^4$ | $10^6$ | $10^8$ | $10^9$ |
|---|---|---|---|---|---|---|---|
| Chinese units: | 十 | 百 | 千 | 万 | ∅ | 亿 | ∅ |
| English units: | ∅ | hundred | thousand | ∅ | million | ∅ | billion |

Table 1 Distributional contrast between English and Chinese numerical units

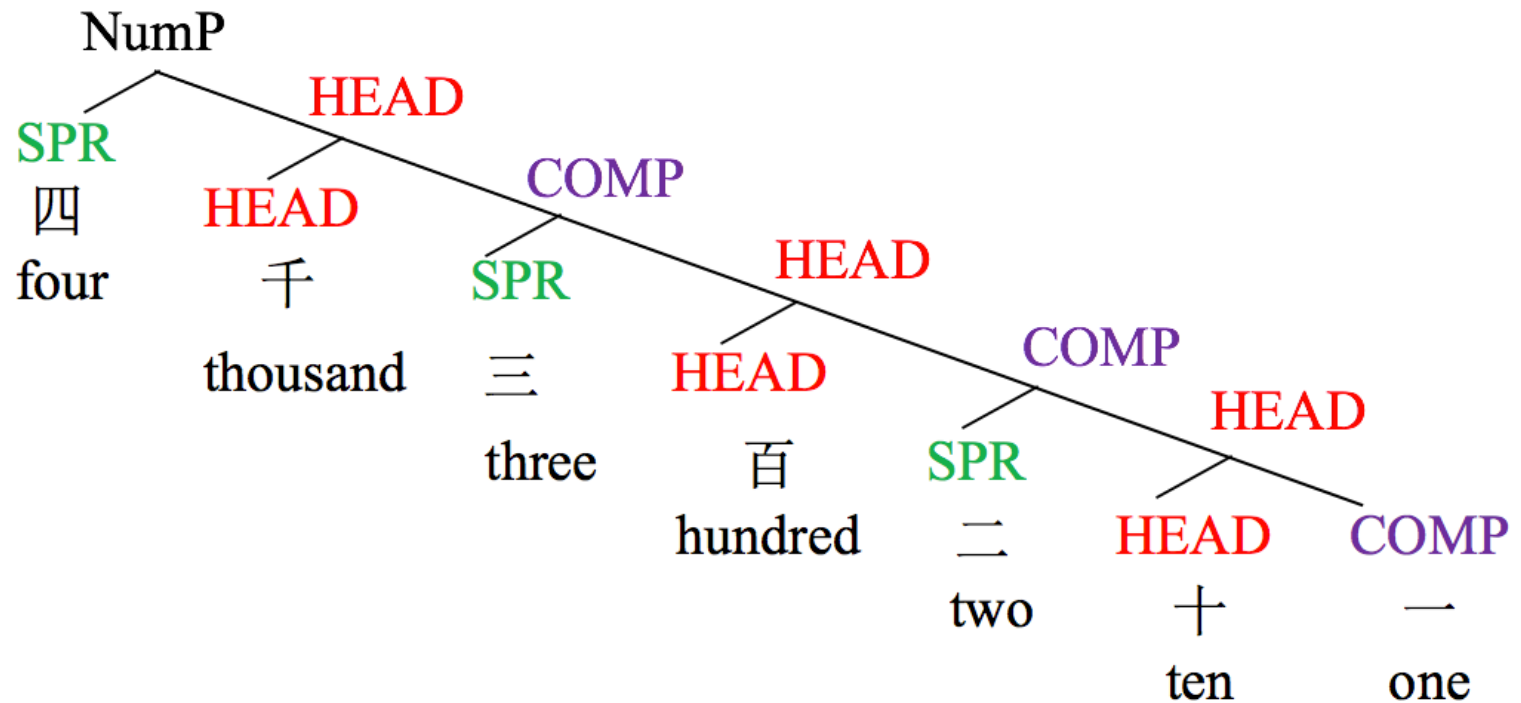# 2.1 Syntax: Regular numerical expressions



Fig. 1 Tree diagram of "4,321 / four thousand three hundred twenty one"

# 2.1 Syntax: Regular numerical expressions

- Constraints on two groups of numerical units with different SPR values.

  - The rigid group, including 十 (ten), 百 (hundred) and 千 (thousand), could only take a digit from 1 to 9 as their specifiers,

  - The free group, including 万 (ten-thousand) and 亿 (hundred-million) can take a saturated numerical expression.
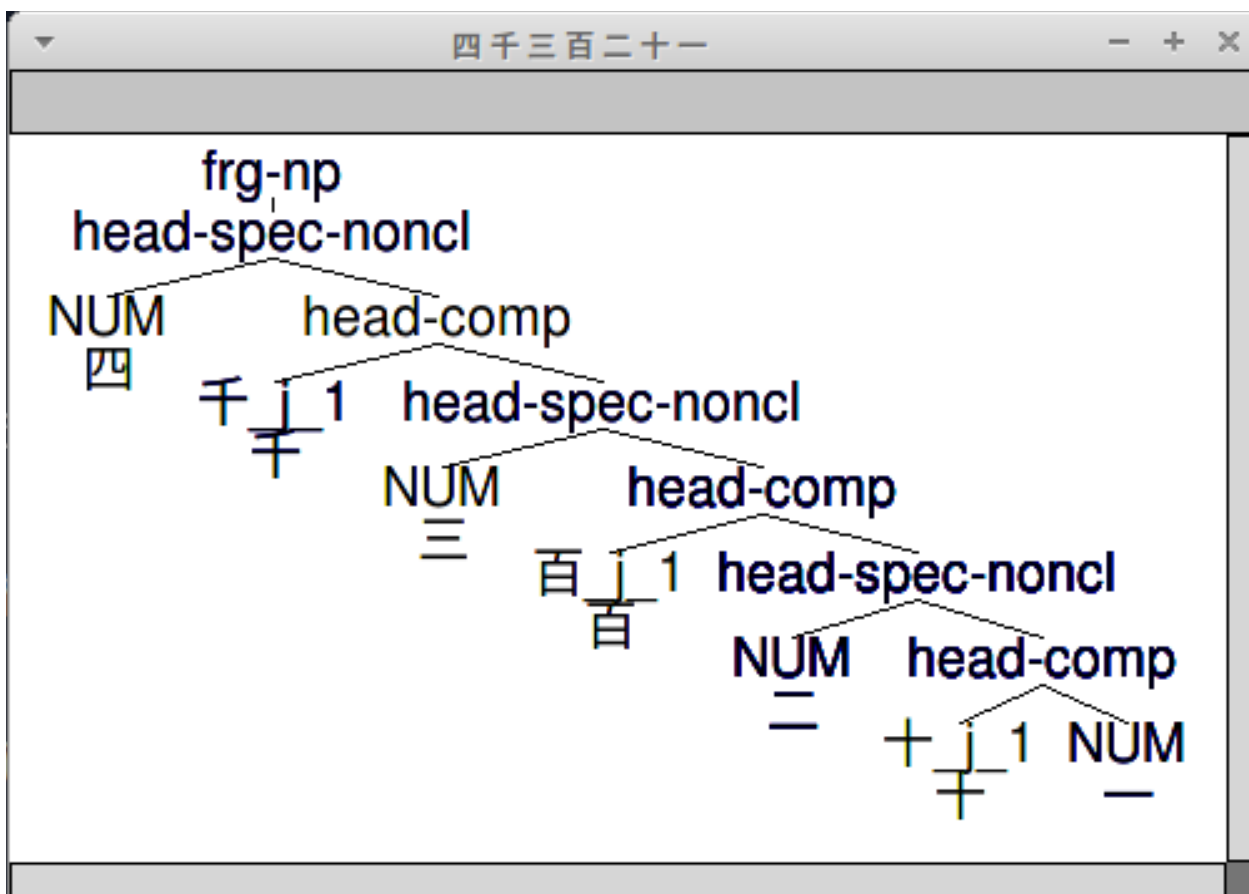
# 2.1 Syntax: Regular numerical expressions



Fig. 2 Parse result of "4,321 / four thousand three hundred twenty one"

# 2.1 Syntax: Regular numerical expressions

More constraints on regular numerals: A numerical unit as head must be IMMEDIATELY lower than that on its left and IMMEDIATELY larger than that on its right (if there is any) on the unit scale illustrated in table 1, e.g.,

a. 二 [千 [一 百] ]

two thousand one hundred

"2,100 / two thousand one hundred"

b. * 一 [百 [二 千] ]

one hundred two thousand
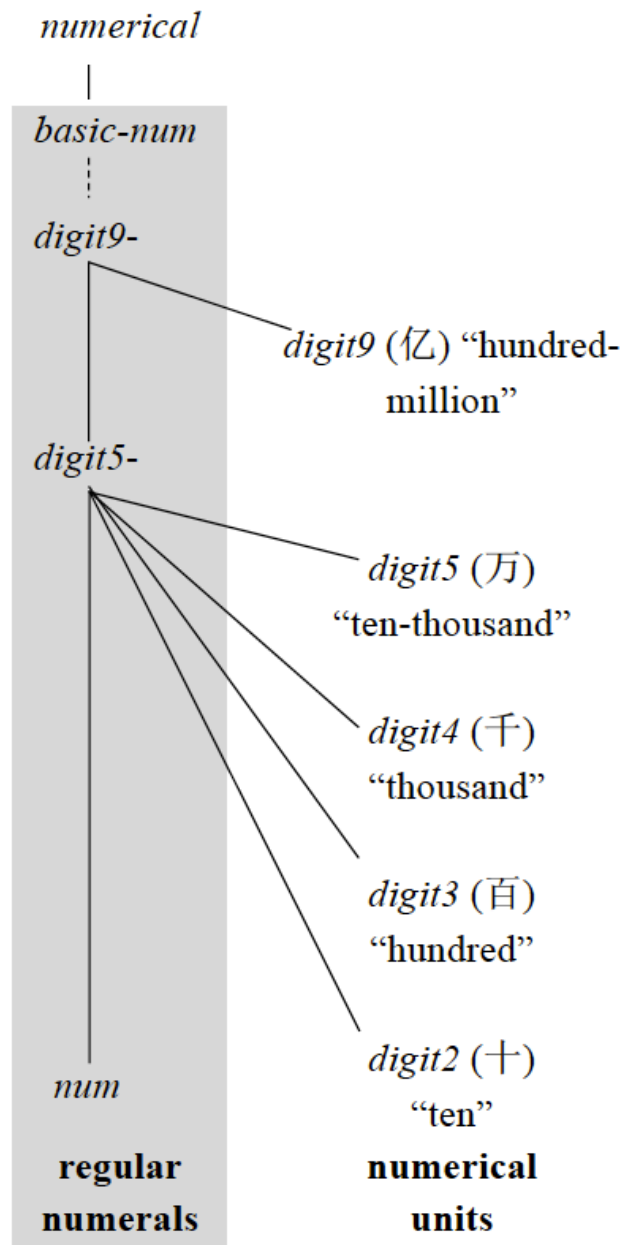
c. * 一 [千 [一 十] ]

one thousand one ten

Fig. 3 Type hierarchy of regular numerical expressions in Mandarin Chinese

# 2.1 Syntax: Regular numerical expressions

**Example definition of 亿 (hundred-million) :**

a_ap_unit-hundred-million_le  :=   basic-numadj-lex & norm−
                                            numadj-lex &

   [ SYNSEM [ LOCAL.CAT [ HEAD digit9,

                VAL [ SPR < [ LOCAL.CAT [ HEAD digit9-,

                         VAL saturated ]]  >,

    COMPS < [ LOCAL.CAT.HEAD digit5 ] > ] ] ] ].

# 2.2 Syntax: 零 (zero) lexical types: distribution

| Phase# | Phase3 | | Phase2 | | | | Phase1 | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Numerical units | | hundred-million | | | | ten-thousand | thousand | hundred | ten | num |
| digit# | … | digit9 | digit8 | digit7 | digit6 | digit5 | digit4 | digit3 | digit2 | digit1 |
| e.g. 1 | | | | 3 | 0 | 0 | 2 | 0 | 0 | 0 |
| e.g. 2 | | 6 | 0 | 5 | 0 | 4 | 3 | 0 | 0 | 2 |

Table 2 Four-digit division numerical system in Mandarin Chinese

**Example 2 in table 2:**

六　　亿　　　　零　五　百　　零　四　万　　三　千　　零　二

six hundred-million zero five hundred zero four ten-thousand three thousand zero two

"605,043,002 / six hundred five million forty-three thousand two"

# 2.2 Syntax: 零 (zero) lexical types: descriptions

- Treat 零 (zero) as a special numerical unit with an obligatory COMP and an empty SPR;

- Build 零 (zero) lexical types and related special numerical units such as 万 (ten-thousand);

- Implement constraints on the distribution of 零 (zero) lexical types.

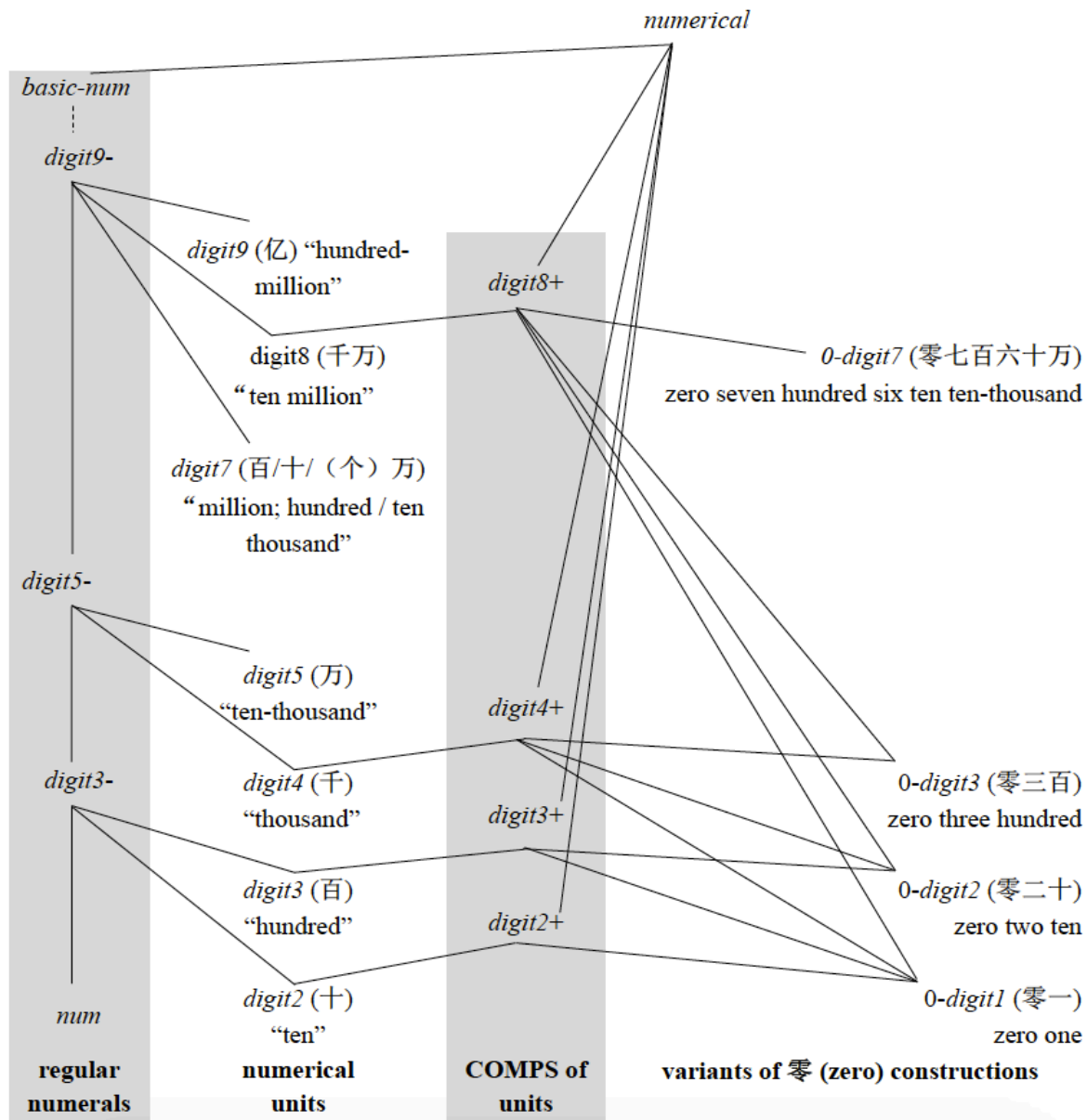- Therefore, a more comprehensive type hierarchy of numerical expressions in Mandarin Chinese …

Fig. 4 Type hierarchy of numerical expressions in Mandarin Chinese

# 2.3 Syntax: "irregular" numerical expressions

- **Syntactic**: e.g., one special type of numerical unit 十 (ten), which, unlike the rest units, takes an optional SPR 一 (one).

- **Stylistic and semantic**: e.g., informal expression 二百五 (two hundred five; "250 / two hundred fifty")

- **Dialectal**: e.g., 二亿一百万 (two hundred-million one hundred ten-thousand; "201,000,000 / two hundred and one million") without 零 (zero) in Taiwan China whereas 零 (zero) is essential in Mainland China.
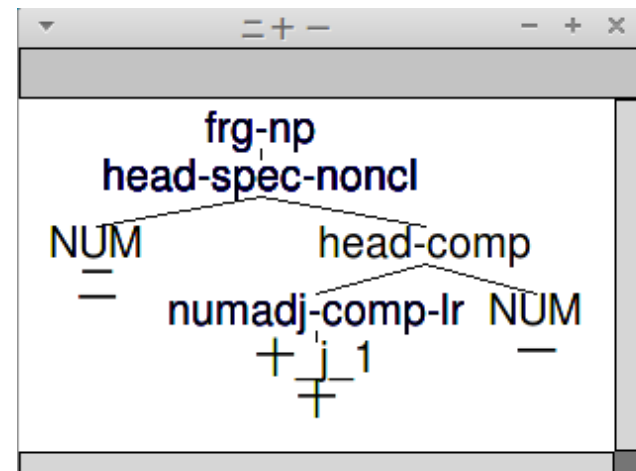
# 3. Semantics of numerical expressions

- Based on Smith (1999)
- Semantics of a numerical expression is a concatenation of multiplication (SPR-HEAD) and addition (HEAD-COMP) relations.
- E.g.,

二　十　一

two  ten  one

"21 / twenty one"

# 3. Semantics: example



Fig. 5  Indexed MRS of 二十一 (two ten one; "twenty one")

# 4. Proficiency check with test suits

- Build two test suites of grammatical and ungrammatical numerical expressions respectively
  - The grammatical one includes 23 numerical expressions of 9 types;
  - The ungrammatical one includes 25 numerical expressions of 7 types.
- The grammar now parses all grammatical sentences and blocks all ungrammatical ones.

# 4. Proficiency check with test suits

- Examples in grammatical test suites:
  - All legal distributive positions of 零 (zero) illustrated in Fig.4
  - Regular and special (no SPR) unit 十 and informal numerical expression as illustrated in section 2.3 "irregular" numerical expressions.
  - Saturated SPR before free units, e.g.,

    三　十　二　　万

    three ten two ten-thousand

    "320,000 / three hundred twenty thousand"

# 4. Proficiency check with test suits

- Examples in ungrammatical test suites:
  - A specifier unit is smaller than a complement unit, e.g.,

    * 一 [百 [二 千] ]

    one hundred two thousand

  - A compulsory 零 (zero) is missing, e.g.,

    * 二 亿 一 百 万

    two hundred-million one hundred ten-thousand

    "201,000,000 two hundred one million"

  - Ungrammatical 零 (zero) distributions, e.g., between adjacent units

# 5. Conclusion

- The current grammar

  - both syntactically and semantically parses regular and some irregular numerical expressions in Mandarin Chinese;

  - reduces redundant parse and generation results with more exact and intuitive constraints.

- Documentation is on the way and comments are welcome …

# References

Pollard, C. J. & I. A. Sag. (1994). *Head-Driven Phrase Structure Grammar*. Chicago: The University of Chicago press.

Guo, R. (2002). Studies of Parts of Speech of Contemporary Chinese. Beijing: Commercial Press.

Lv, S. (2002). Collections of Lv Shuxiang (vol. 5: Eight Hundred Words of the Contemporary Chinese). Shenyang: Liaoning Education Press.

Smith, J. D. (1999). English number names in HPSG. In G. Webelhuth, J.-P. Koenig, & A. Kathol (Eds.), *Lexical and Constructional Aspects of Linguistic Explanation* (pp. 145-160). Stanford: CSLI Publications.