汉语自然数表达形式特征分析

詹卫东
北京大学中文系

数字表达的范围

1～9999 或 1～9999,9999		1～9999		1～9999
X	亿	Y	万	Z

12位数： 1～9999 9999 9999

九千九百九十九 **亿** 九千九百九十九 **万** 九千九百九十九

16位数： 1 ～ 9999 9999 9999 9999

九千九百九十九万九千九百九十九 **亿** 九千九百九十九 **万** 九千九百九十九

16位数以上：

万万亿 …… 十万万亿 …… 百万万亿 …… 千万万亿

万万万亿 …… 十万万万亿 …… 百万万万亿 …… 千万万万亿

万万万万亿 …… 十万万万万亿 …… 百万万万万亿 …… 千万万万万亿

1） 系数：一 二（两）三 四 五 六 七 八 九

2） 位数： 十 百 千 万 亿 （兆）

3） 基本结构：系 + 位 e.g. 八十

4） 组合结构：（系 + 位）+（系 + 位）+ ……

最末一位的“位数”可以省略，比如“一百八”。省略了位数的系位构造，只能后置，不能前置。

5）系位组合中前后“位”之间有数值从大到小的顺序关系。

6）如果不是紧邻的位数，则在两个系位构造（“系 + 位”）之间要用“零”来占位分隔。

7）“十”作为位数使用，前面的系数是“一”时，“一”可以省略，比如：“十八”，指的是“一十八”。其他的位数词“百，千，万，亿”没有这种用法（*百八，*千八）。

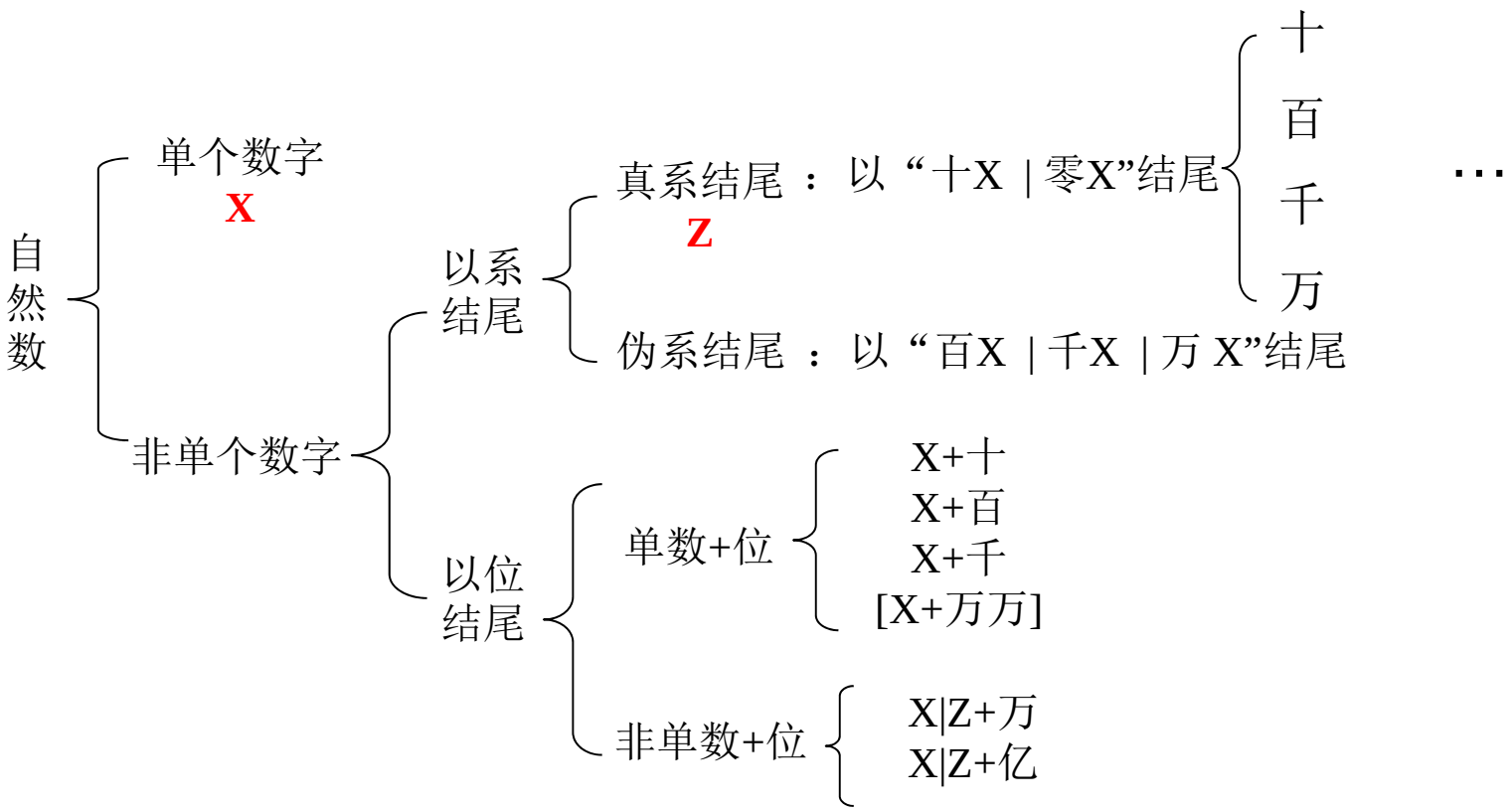
8）“万、亿”跟“十、百、千”不同。

“万、亿”这两个位数前面的系数可以是系位组合；“十、百、千”前面的系数只能是从“一”到“九”的系数。

“万”前面的系数取值在1到9999之间；

“亿”前面的系数取值在1到99999999之间。

9）“万万”可以看做是由基本位数组合而成的位数，跟“千万、百万、十万”不同。因为“万万”的后面还可以出现“千万”，比如“四万万五千万”，如果把“四万万”分析为：四万 + 万，这样就会出现连续两个“万”位数的构造，即（四万 + 万）+（五千 + 万）。


自然数
单个数字
X
非单个数字
以系结尾
真系结尾 ： 以 “十X | 零X”结尾
Z
十
百
千
万
…
伪系结尾 ： 以 “百X | 千X | 万 X”结尾
以位结尾
单数+位
X+十
X+百
X+千
[X+万万]
非单数+位
X|Z+万
X|Z+亿

构造CFG规则时应注意的问题举例

- “两”跟“二”的分布差异

	__十	__百|千|万|亿	十|百|千|万__	亿__	零__
两	-	+	-	-	-
二	+	+	+	-	+

- “十五”跟“一十五”的分布差异

	百__	千|万|亿 零 __	__万|亿
十五	-	?	+
一十五	+	+	+

构造CFG规则时应注意的问题举例(续)

- “三百”跟“三百二十一”的分布差异

  ? 三百 万 三千 —— 三百二十一 万 三千

- “三百二”跟“三百二十”的分布差异

  * 三百二 万 —— 三百二十 万

三个需要考虑的因素

- 数值大小
- 开头
- 结尾

作业要求

1. 提交CFG规则文件（.txt 纯文本文件格式）
2. 提交实验报告（需包含下列内容）：

（1） 规则编写过程说明。
（2） CFG规则的条数，非终结符的个数，
每一个非终结符的分布特征。
（3） 是否存在有关联的非终结符，即分布上有相同的位置，同时又有不同的位置。
（4） 测试报告：

- 测试了多少正例，多少反例。（给出规则制定者存疑的形式。比如“五十五万六”）
- 对于正例，分析结果是否有多个（歧义）的情况。如果有，原因是什么。
- 总的正确率或错误率。