

从利用语言直觉到计算模型

汉语自动分词

杨沐昀

哈工大教育部-微软语言语音重点实验室

MOE-MS Joint Key Lab of NLP and Speech (HIT)

基于N元语法的切分排歧

- * 基于HMM的分词词性标注一体化模型

$$P(\text{POS} | \text{Sentence}) = P(\text{POS} | W)$$

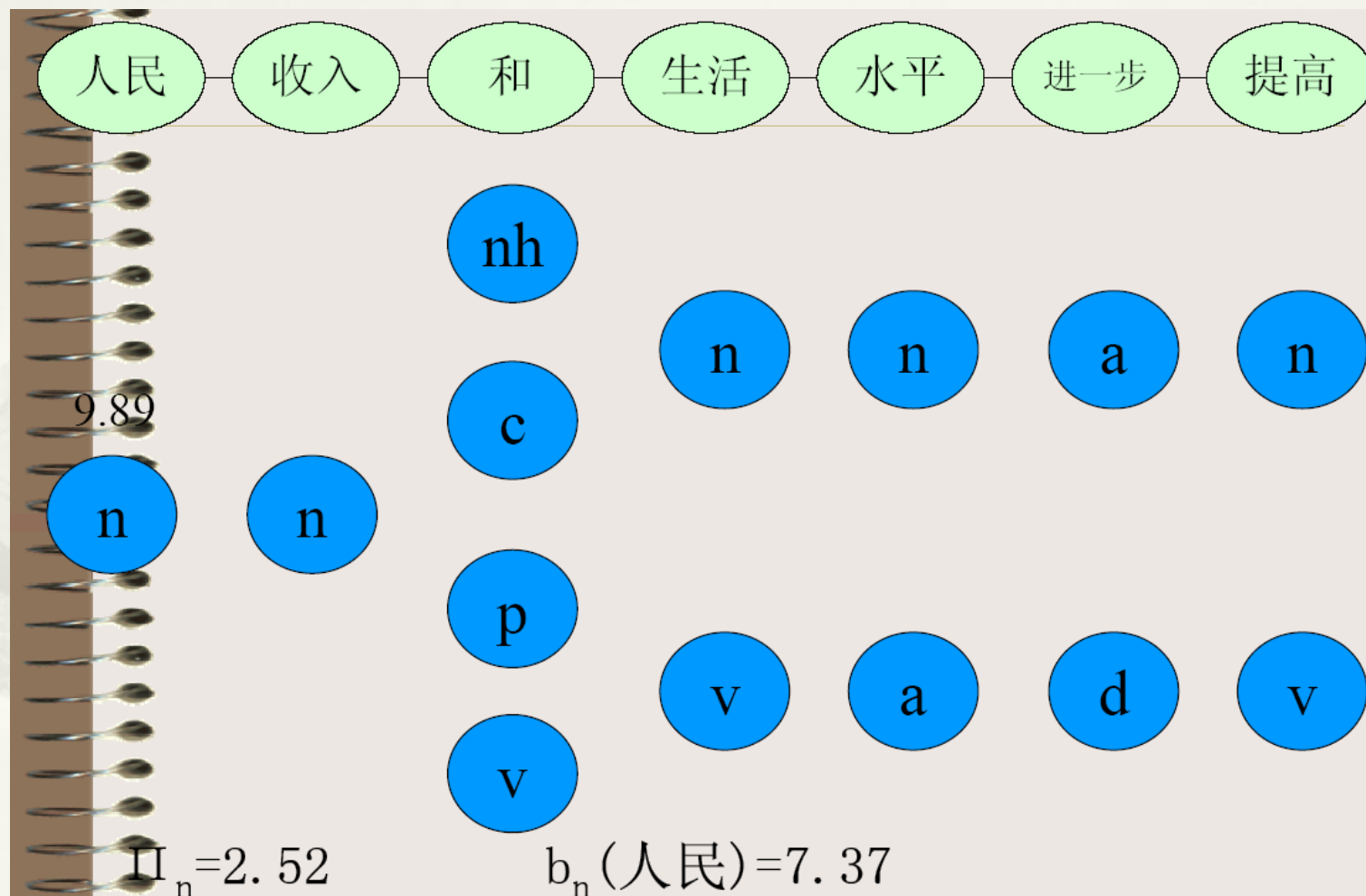
$$\Rightarrow P(W | \text{POS}) \cdot P(\text{POS})$$

*W为分词序列

- * 采用二元文法计算上式

$$P(W_1 | \text{POS}_1) \cdot P(\text{POS}_1 | S_{\text{begin}}) \cdot P(W_2 | \text{POS}_2) \cdot P(\text{POS}_2 | \text{POS}_1) \cdot \dots \cdot P(W_n | \text{POS}_n) \cdot P(\text{POS}_n | \text{POS}_{n-1})$$

Viterbi搜索——例子



人民 收入 和 生活 水平 进一步 提高

nh

n

n

a

n

9.89

20.02

c

n

n

p

v

a

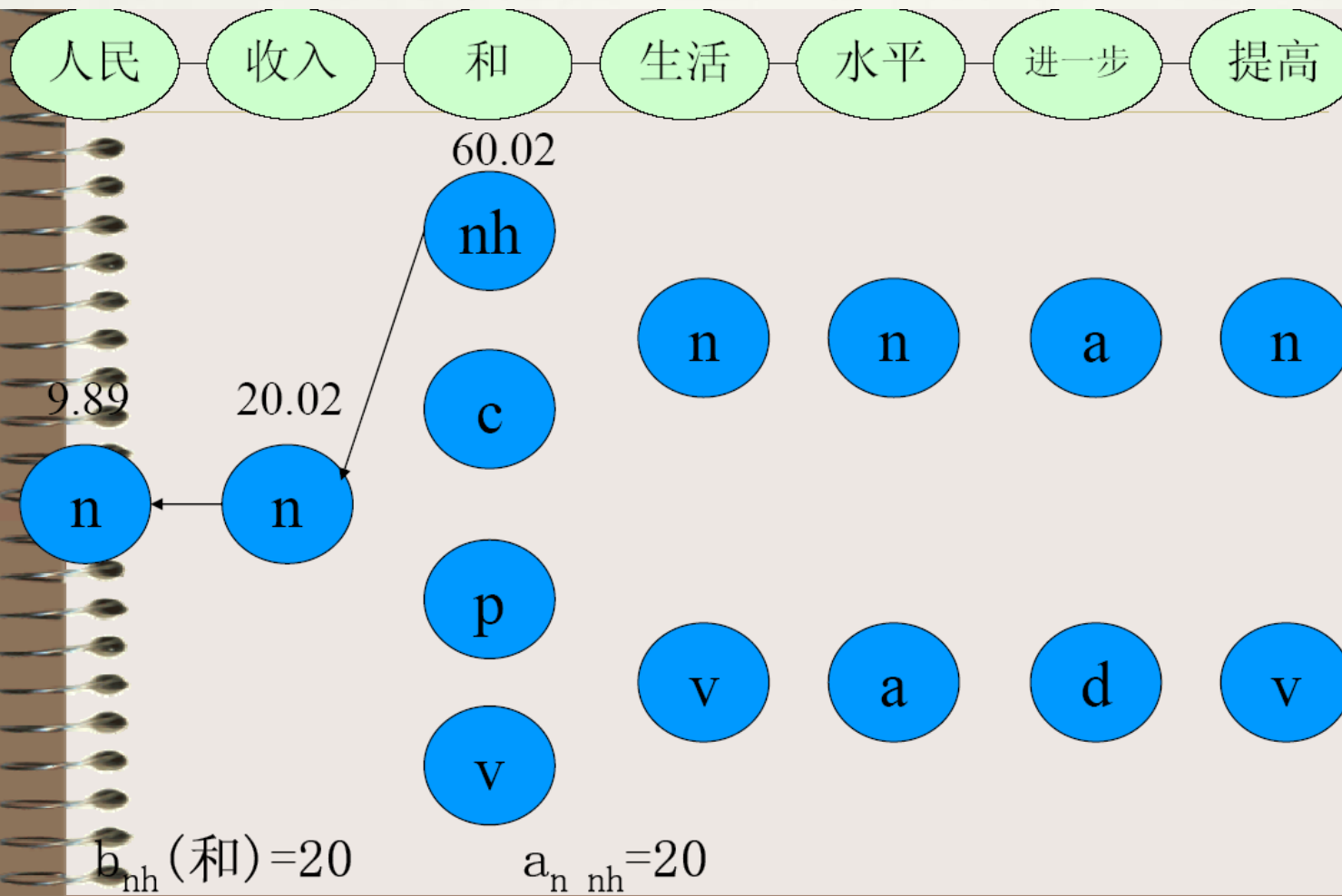
d

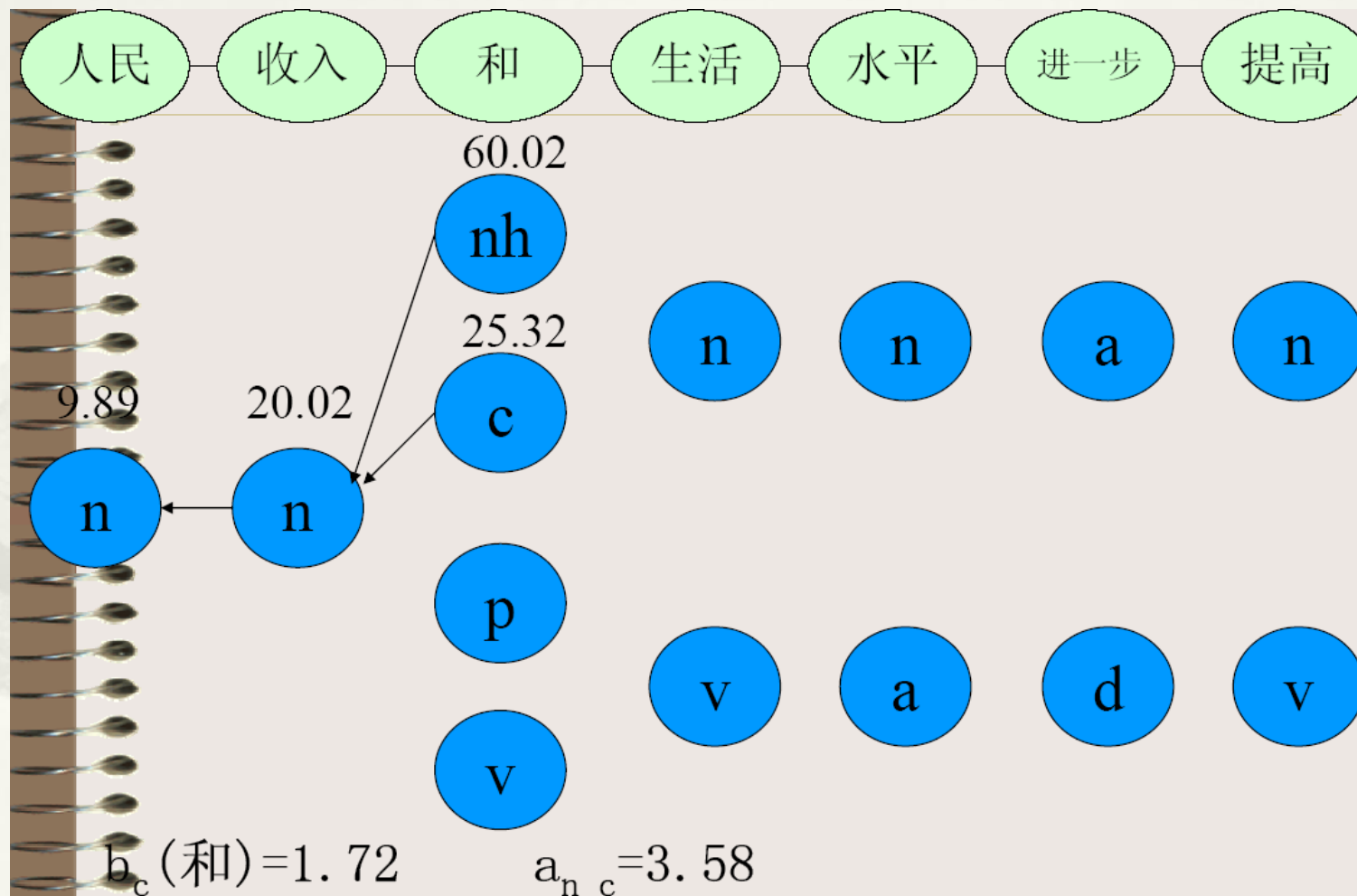
v

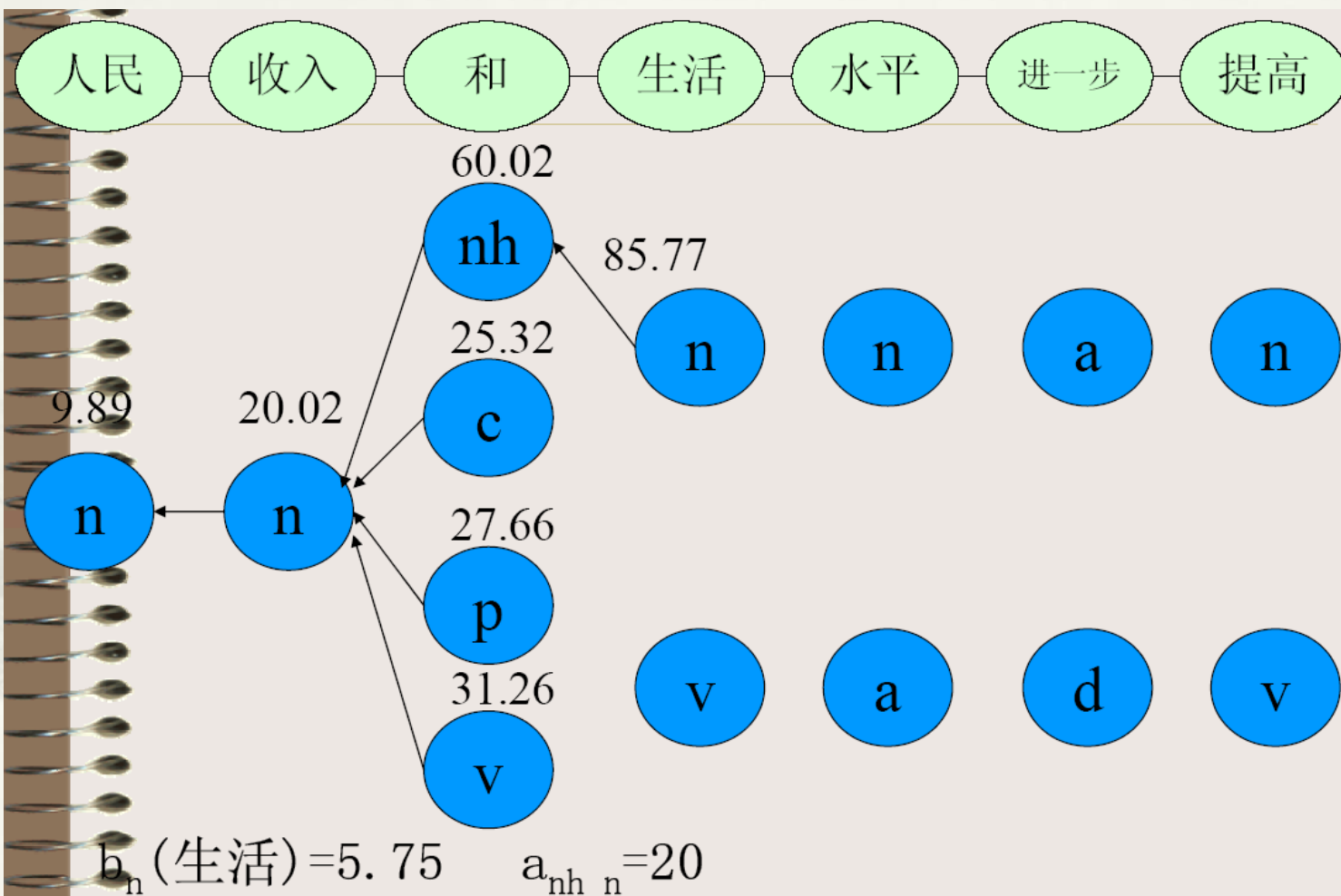
v

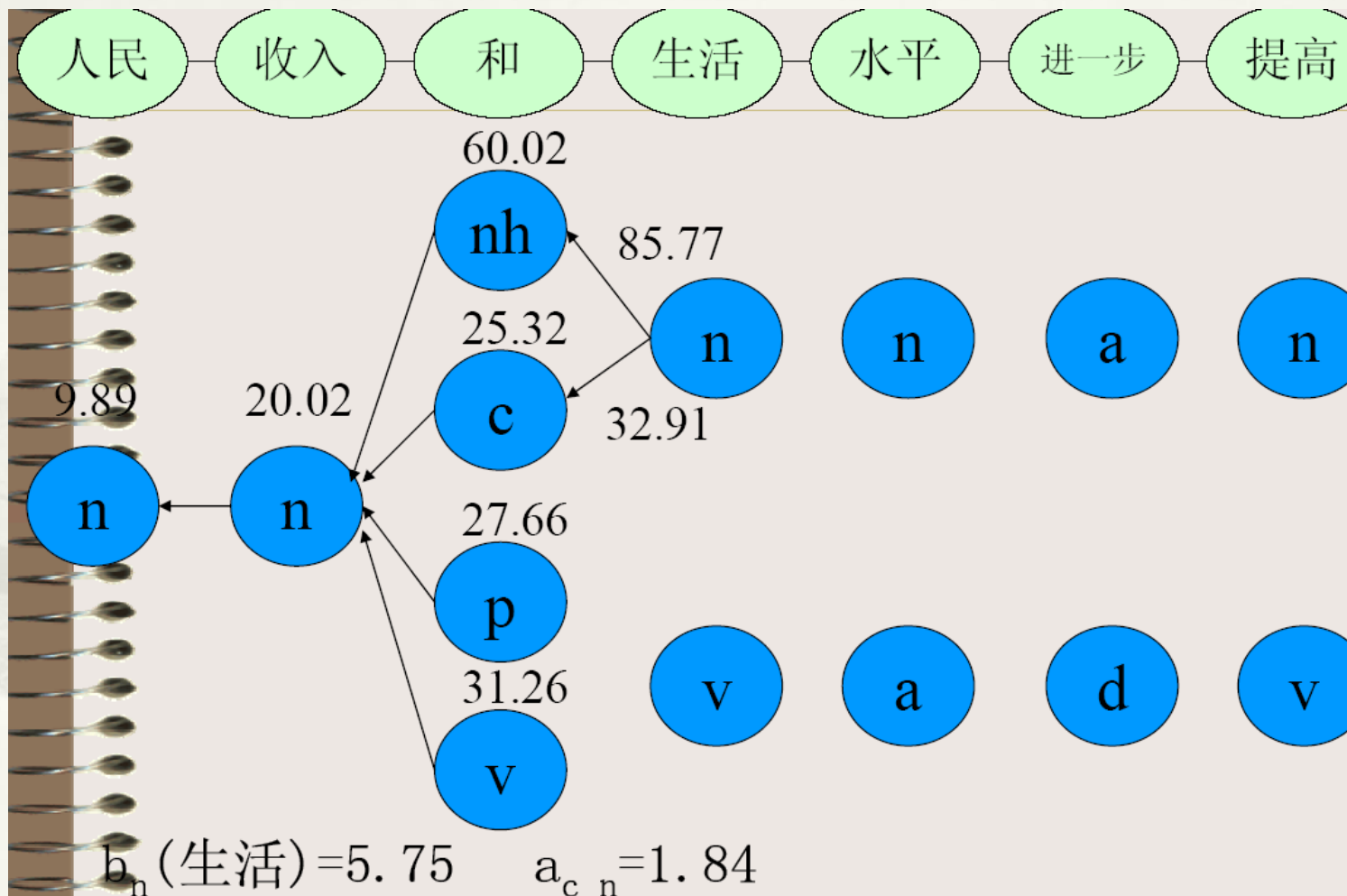
$b_n(\text{收入}) = 6.98$

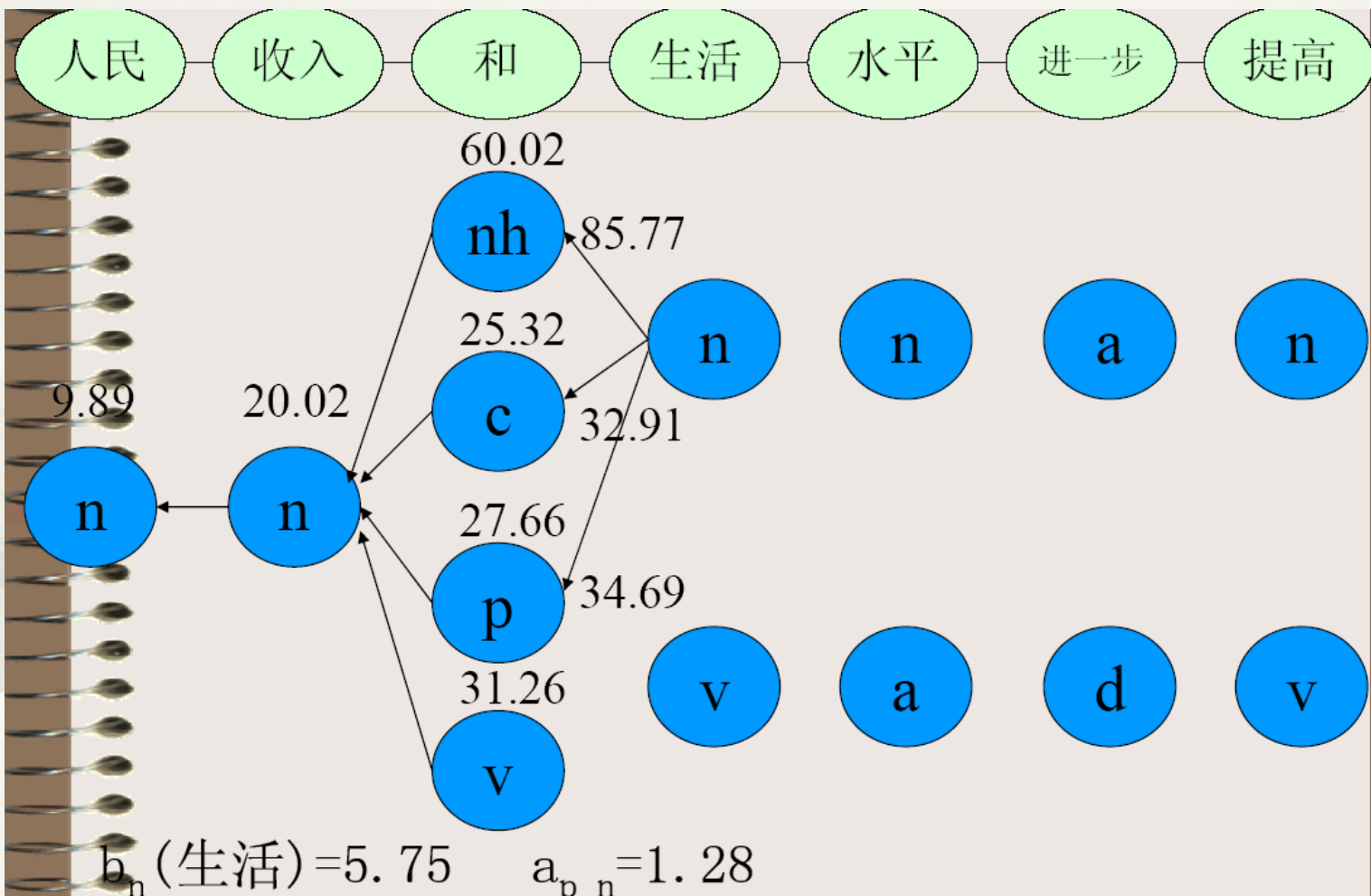
$a_{nn} = 3.15$

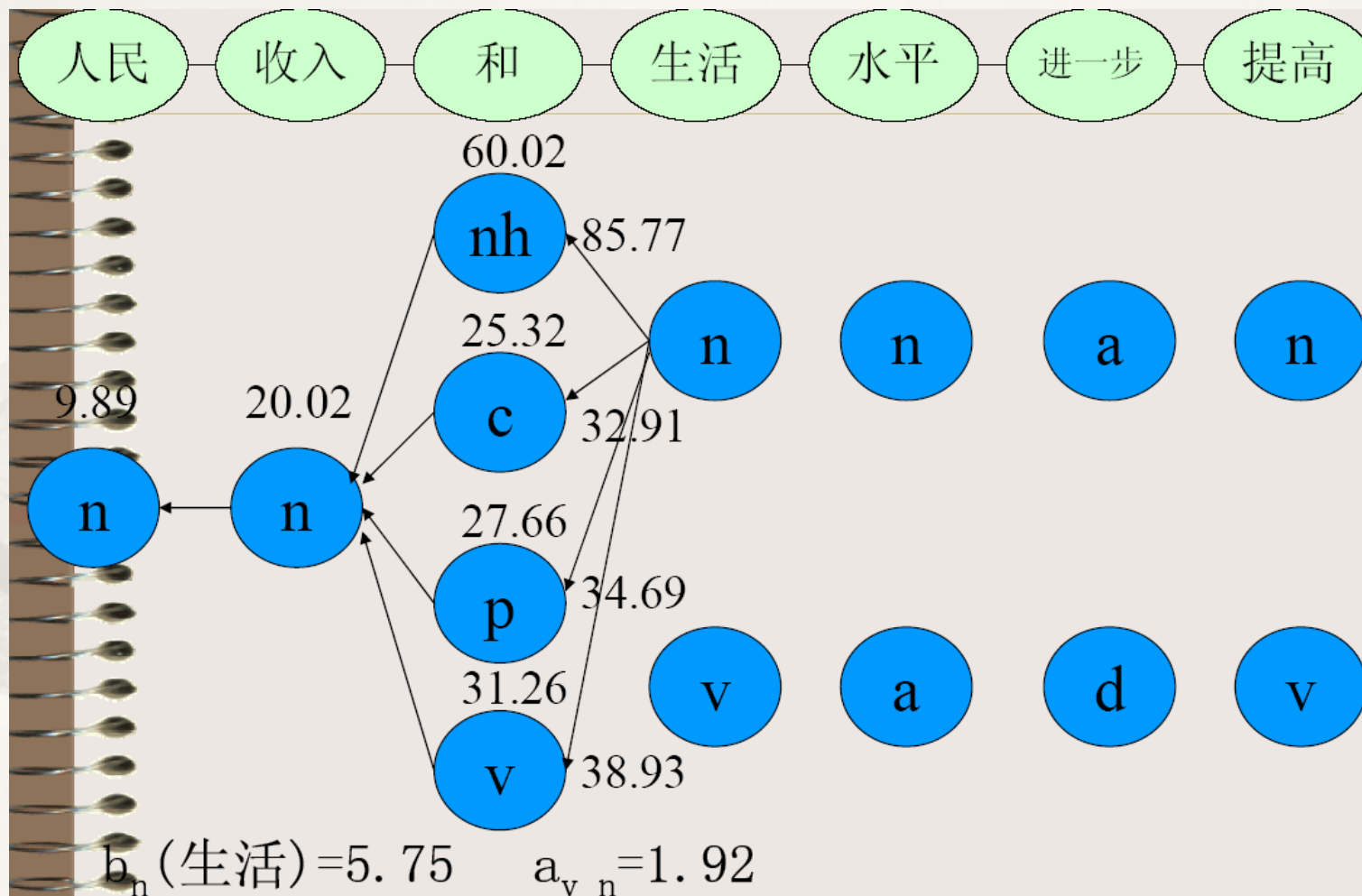


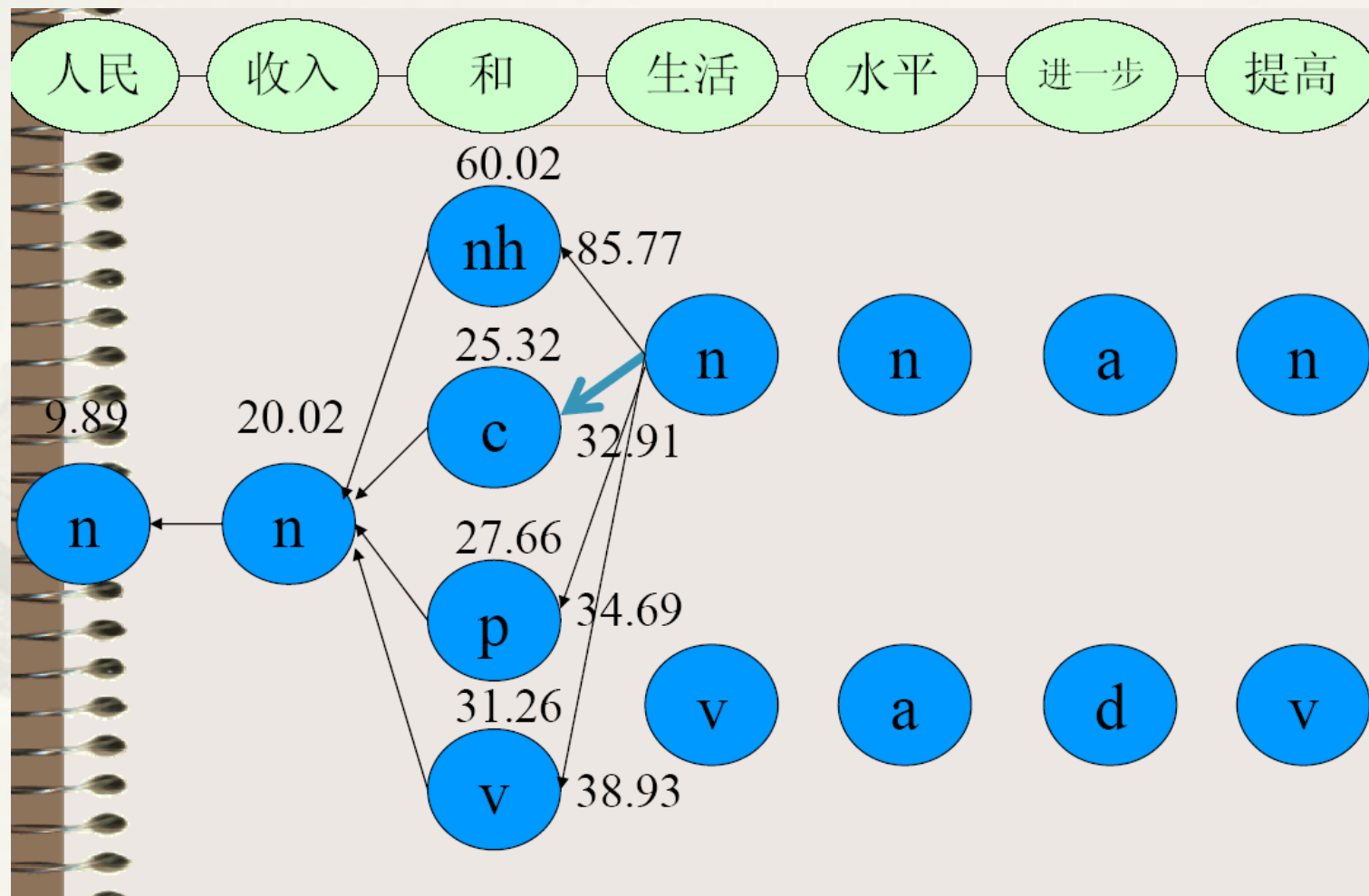


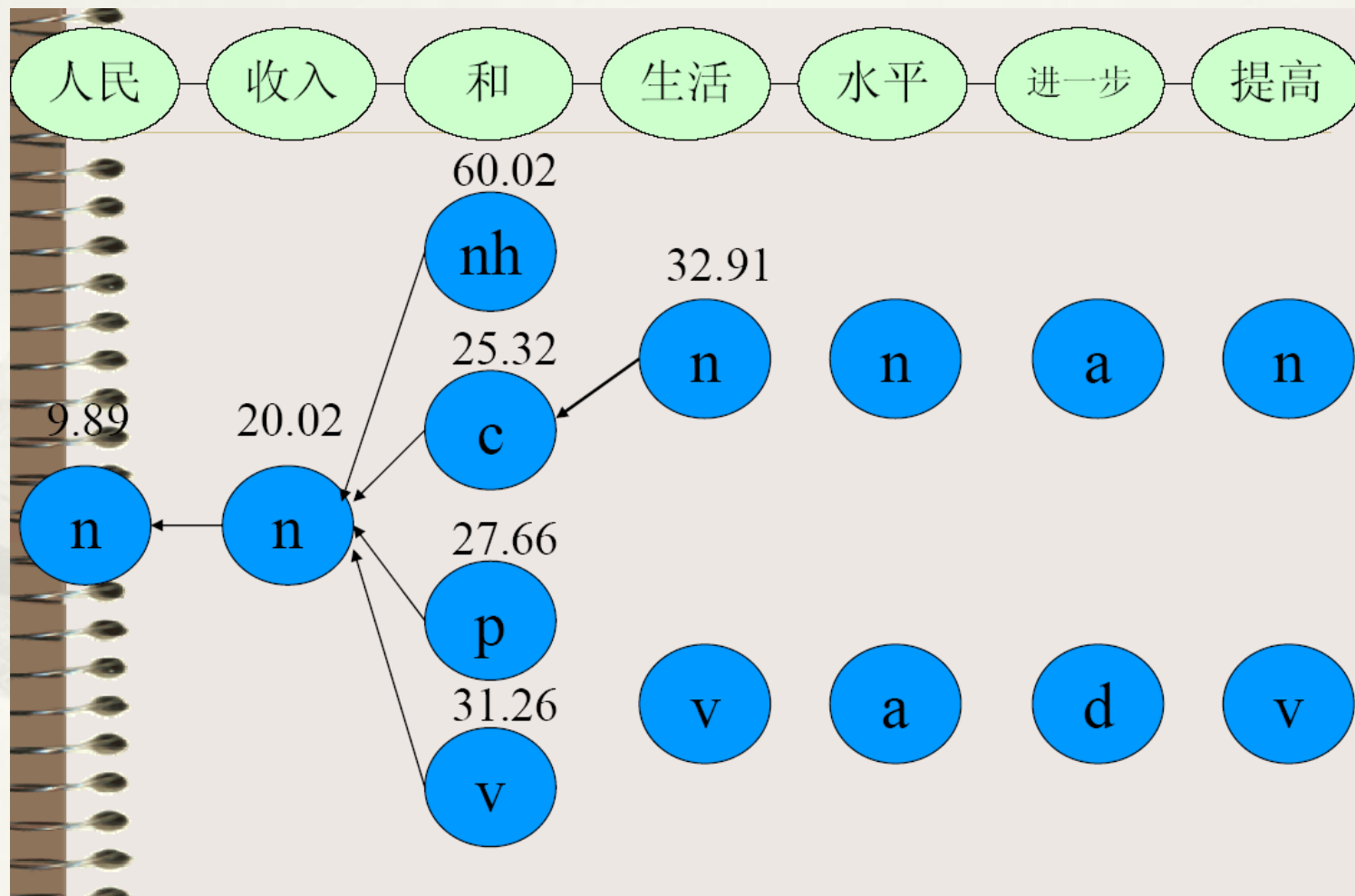


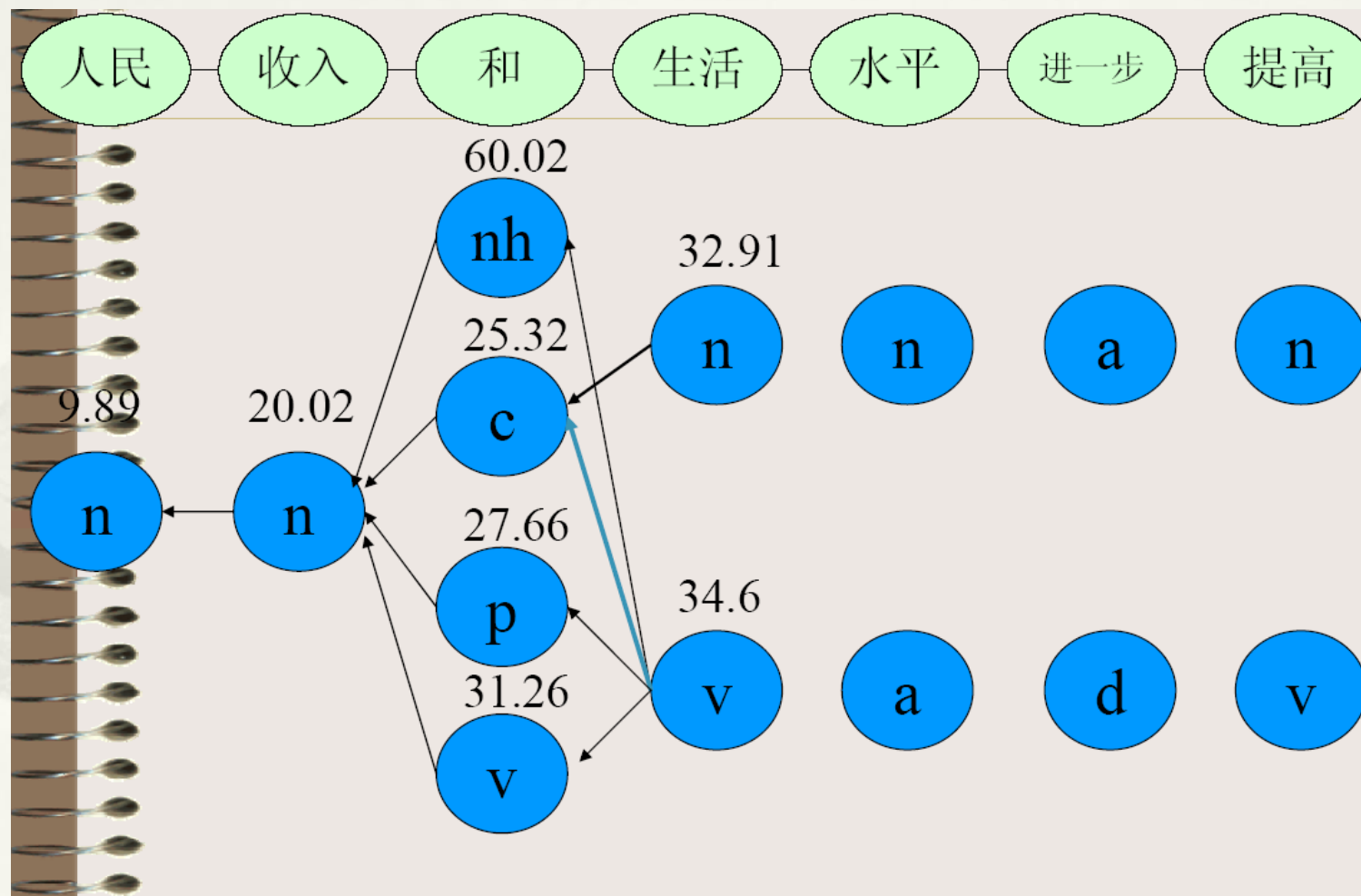


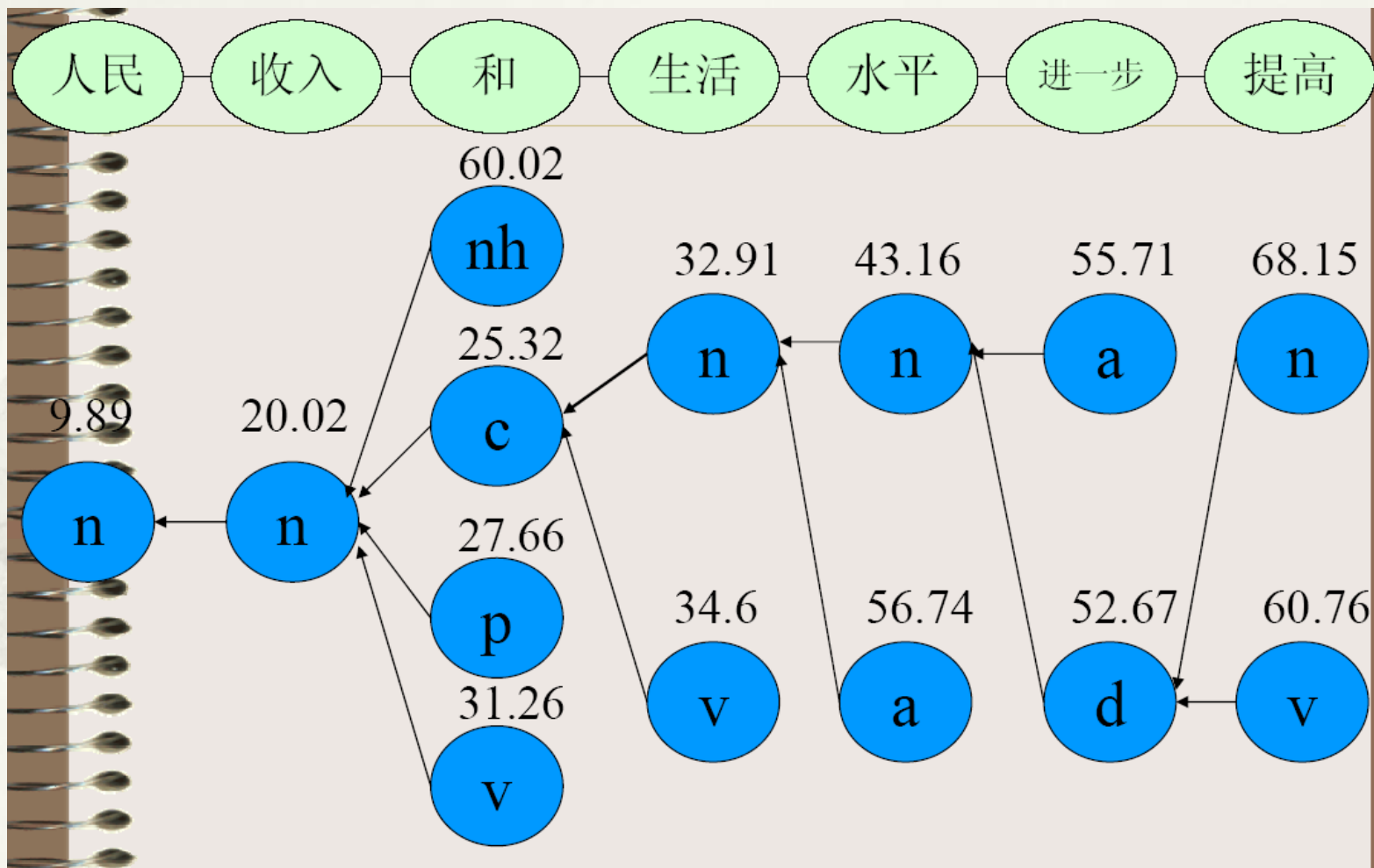


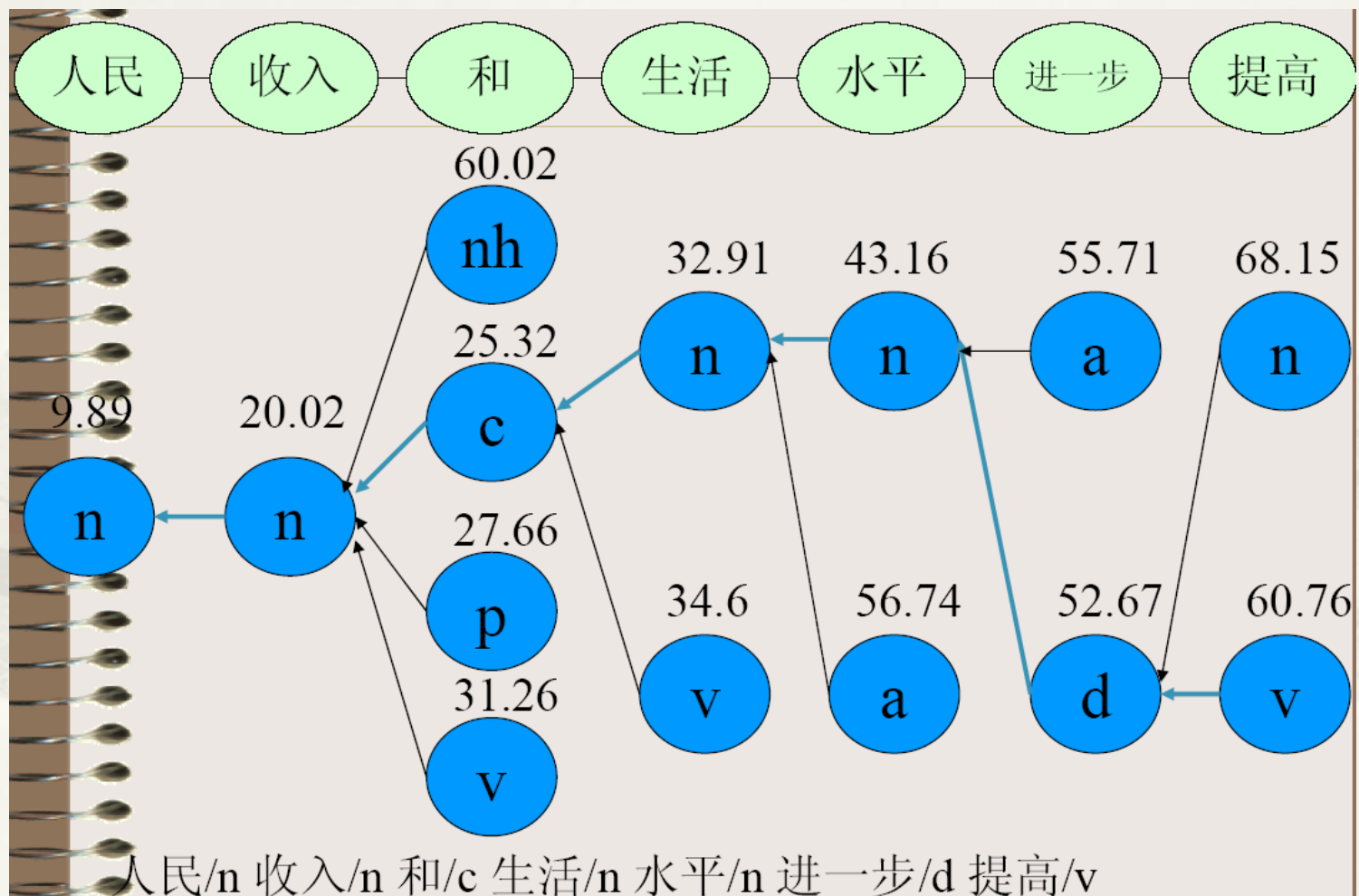












中文未登录词识别

The background of the slide features a large, semi-circular graphic that resembles an open traditional Chinese folding fan. The fan's surface is covered with a faint, monochromatic landscape painting in a traditional Chinese style, depicting mountains, trees, and a winding path. The fan is positioned centrally, with its edges curving towards the bottom corners of the slide. The overall color palette is a soft, muted green, giving the slide a classic and scholarly appearance.

未登录词的类型

- * 命名实体（Named Entity）

- * 汉语人名：李素丽 老张 李四 王二麻子
- * 汉语地名：定福庄 白沟 三义庙 韩村 河马甸
- * 翻译人名：乔治·布什 叶利钦 包法利夫人
- * 翻译地名：阿尔卑斯山 新奥尔良 约克郡
- * 机构名：方正公司 联想集团 国际卫生组织外贸部

- * 数字、日期词、货币等

- * 商标字号：非常可乐 乐凯 波导 杉杉 同仁堂

- * 专业术语：万维网 主机板 模态逻辑 贝叶斯算法

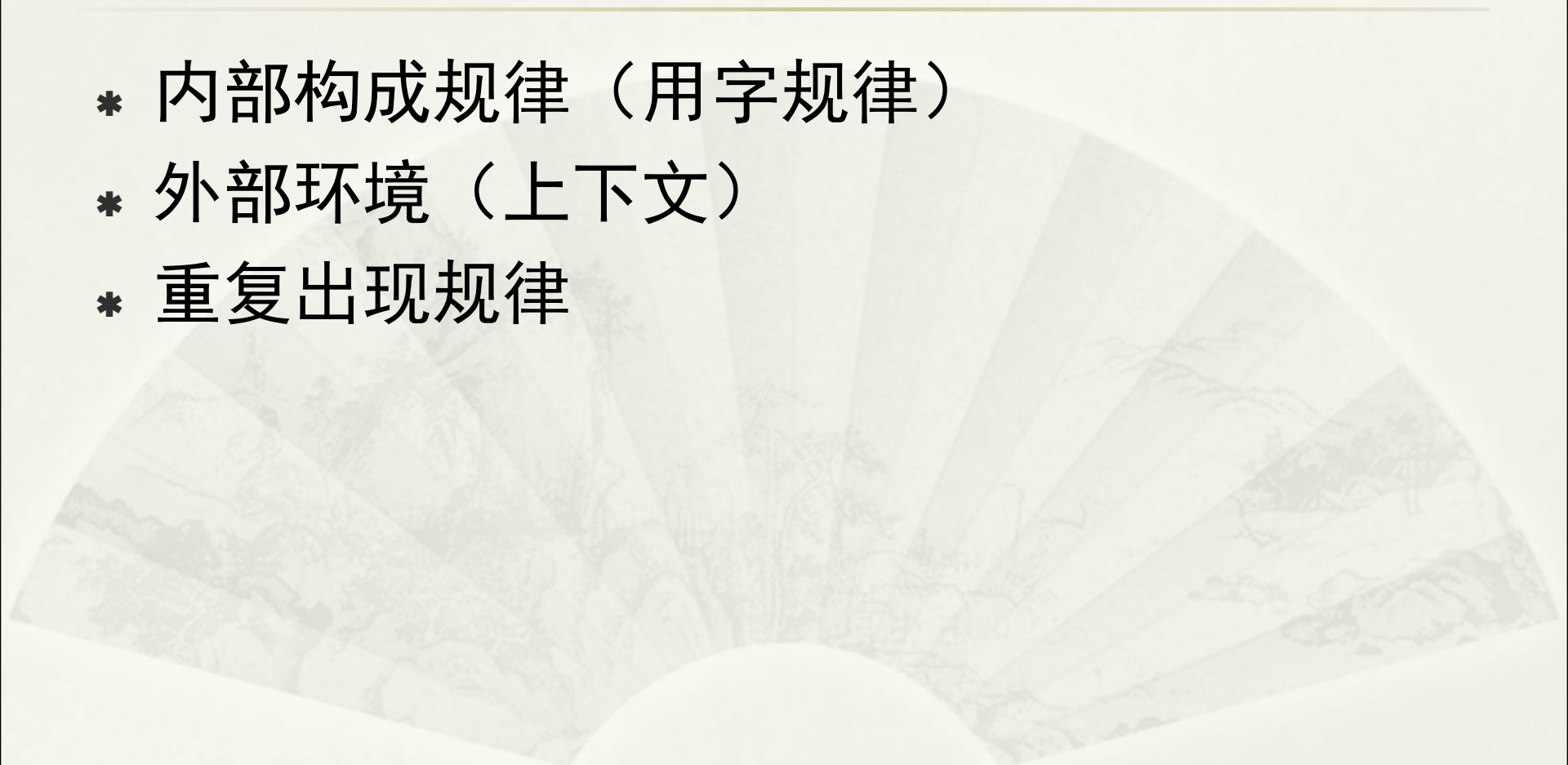
- * 缩略语：三个代表 五讲四美 打假扫黄 打非计生办

- * 新词语：卡拉OK 波波族 美刀 港刀

未登录词识别的困难

- * 未定义词没有明确边界
- * 未定义词的构成单元（汉字）本身都可以独立成词

未登录词识别的依据

- * 内部构成规律（用字规律）
 - * 外部环境（上下文）
 - * 重复出现规律
- 

未登录词识别的研究进展

- * 很成熟：
 - * 数字、日期、货币词
- * 较成熟
 - * 中国人名、译名
 - * 中国地名
- * 较困难
 - * 商标字号
 - * 机构名
- * 很困难
 - * 专业术语
 - * 缩略语
 - * 新词语

中国人名的内部构成规律

- * 在汉语的未定义词中，中国人名是规律性最强，也是最容易识别的一类；
- * 中国人名一般由以下部分组合而成：
 - * 姓：张、王、李、刘、诸葛、西门、范徐丽泰
 - * 名：李素丽，张华平，王杰、诸葛亮
 - * 前缀：老王，小李
 - * 后缀：王老，赵总
- * 中国人名各组成部分用字比较有规律

中国人名的内部构成规律

- * 台湾出版的《中国姓氏集》收集姓氏5544个，其中，单姓3410个，复姓1990个，3字姓144个。
- * 中国目前仍使用的姓氏共737个，其中，单姓729个，复姓8个。
- * 根据我们收集的300万个人名统计：姓氏：974个，其中，单姓952个，复姓23个，300万人名中出现汉字4064个。

中国人名的内部构成规律

- * 中国人名各组成部分的组合规律
 - * 姓 + 名
 - * 姓
 - * 名
 - * 前缀 + 姓
 - * 姓 + 后缀
 - * 姓 + 姓 + 名（海外已婚妇女）

中国人名的上下文构成规律

* 身份词：

- * 前：工人、教师、影星、犯人
- * 后：先生、同志
- * 前后：女士、教授、经理、小姐、总理

* 地名或机构名：

- * 前：静海县大丘庄禹作敏

* 的字结构

- * 前：年过七旬的王贵芝

* 动作词

- * 前：批评，逮捕，选举
- * 后：说，表示，吃，结婚

中国人名识别的难点

- * 一些高频姓名用字在非姓名中也是高频字
 - * 姓氏：于，马，黄，张，向，常，高
 - * 名字：周鹏和同学，周鹏和同学
- * 人名内部相互成词，指姓与名、名与名之间本身就是一个已经被收录的词
 - * [王国]维、[高峰]、[汪洋]、张[朝阳]
- * 人名与其上下文组合成词
 - * 这里[有关]天培的壮烈；
 - * 费孝通向人大常委会提交书面报告
- * 人名地名冲突: 河北省刘庄

中文姓名识别方法

* 中文姓名识别方法

- * 姓名库匹配，以姓作为触发信息，寻找潜在的名字
- * 计算潜在姓名的概率估值及相应姓氏的姓名阈值，根据姓名概率评价函数和修饰规则对潜在的姓名进行筛选。

中国地名的识别

* 困难

- * 地名数量大，缺乏明确、规范的定义。
《中华人民共和国地名录》（1994）收集88026个，不包括相当一部分街道、胡同、村庄等小地方名称。
- * 真实语料中地名出现情况复杂。如地名简称、地名用词与其它普通词冲突、地名是其它专用名词的一部分，地名长度不一等。

未登录词识别的一般方法

- * 各种不同类型的未登录词识别方法思想大同小异，但实现时各有侧重
- * 各种不同类型的未登录词识别都需要收集大量数据，建立不同的数据模型
- * 常用的方法包括
 - * 规则方法：人工总结或归纳出一些判别规则，并用程序实现
 - * 统计方法：建立统计模型，通过人工标注语料库进行参数训练

将识别问题转化成标注问题

- * 在统计方法中，未登录词识别的一种最通常的做法就是将识别问题转化成标注问题
- * 汉字序列的标注问题可以采用隐马尔科夫模型（HMM）、最大熵（ME）、最大熵马尔科夫模型（MEMM）、条件随机场（CRF）等模型来解决