关于音码方案和形码方案的对比分析

梁金宝

2018 年 3 月 23 日

1. 编码方案选型问题

关于蒙古文编码方案，在制定一种方案或者选定一种方案时，我们应当考虑如下两个因素。

- 方案本身的优缺点

任何一种方案不可能是只有优点而没有缺点。音码方案有它的优点，形码方案也有它的缺点。

- 实施新方案时的社会成本

如果是第一次在多种方案中选定一个方案，那么方案的优点所占权重很重要。很多时候可能直接成为选定理由。但是如果是在考虑用一种方案去替换已经在实施的现行方案，那么需要考虑优缺点以外的社会成本。社会群体、机构的规模越大，实施新方案所付出的社会成本越大。

2. 对比分析两种方案的公平性

蒙古文编码的现状来考虑，现行音码方案确实存一些问题。其中大部分不是方案本身特征造成的。对比的目的是在尝试找解决问题的最佳方案。一个方案最终实施效果的好坏取决于两个方面：

- 方案本身固有的缺陷问题
- 实施方案的方式方法问题

只有对比方案本身的固有问题，才能达到我们的对比目的。否则我们可能会被眼前的音码方案现状所迷惑，被诱导做出过激的变革。

为了完善现行音码方案，最终达到稳定统一的编码体系，我们也做了一份对现行音码体系的修订方案。详细参见另一个文档“关于现行音码方案的最小化修订方案.docx”。

3. 社会需求的多样性

一切方案都不可能是万能的。我们的解决方案是处在一个多种需求同时存在的复杂环境中。因为一种方案在解决一种需求时是完美的，但在解决另一种需求时候他的缺陷会表现出来。比如如下表格：

	词码到词义 Code2mean	词形到词码 Shape2code	搜索 Search	排序 Sort	输入法 IME	用户直觉 sensible intuition	字体实现 Font Implement
Phonetic	★			★	★	★	
Graphetic		★	★				★

※asterisk is mean that There is an advantage.

因为在不同领域每种需求的重要性比重不同，所以不能用星的多少来衡量好坏。这表只是想说明有这些多维因素会影响一种方案的好坏。并且各种技术发达的如今，每种方案的缺点可以通过辅助性工具或措施来弥补。如搜索（形码优势），排序（音码优势），输入法（音码优势）等等。所以我们应当慎重考虑追求纯粹理论上的完美方案。

4. 字体实现

形码方案在字体制作方面的方便性是毋庸置疑的。而现行音码方案在编码统一、或者字体行为方面存在不兼容性问题也是客观事实。但是造成此现象的原因有两个：方案本身的固有缺陷和实施方案时的方法不当。

(1) 现行音码方案模型非常复杂，详细规范不足问题

模型确实复杂。如果仅仅从变形规则角度考虑，蒙古文变形规则本身就比阿拉伯文的复杂很多。现行音码方案确实缺少统一的详细变形规范。并且存在几个互相不兼容的变形规则版本。但是需要理解的是这个不是现行方案本身固有原因造成的，而是没有及时制定相关国家标准造成的。

对此，我们已经正在尝试制定项目标准，并以项目形式实施了推动统一和规范变形规范的工作。目前取得了很好的效果。已经在中国的几家重要信息化企业字体上做到了统一，如蒙科立、德力海、嘎拉图等等。今年年底的时候几乎所有蒙古文网站都会升级到支持此项目规范的字体和输入法上。

(2) 现行音码方容易导致有歧义的文本表记

相比于形码方案，有同形不同码是音码方案的一项弊端。造成歧义文本的元音基本有如下几种：

A.方言差异造成混同元音 O/U/OE/UE

这种比重最大。基本上来自懂蒙古文、但有方言差异的地方。

B.恶意拼凑字形（字母&FVS）

这种比重小，懂或不懂蒙古文。一般出现在只有打印和显示需求的行业。

C.看图录入单词

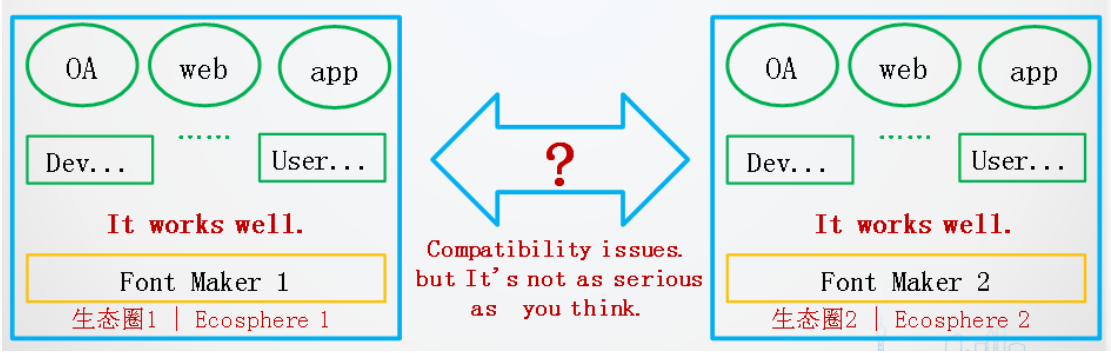
这种比重小，不太蒙古文，按键逻辑来自眼睛看见的单词外形。

D.不可预期的多余 FVS 干扰

这种比重小，人为误操作不可见 FVS 时的顺带产物

此问题到此有多严重？如果仅仅从研究者角度考虑，此问题越研究感觉越严重。其实在实际社会生产环境中问题是客观存在，但是没有那么明显。或者达到必须替换编码方案的程度。它所带来的弊端，远远小于新方案的弊端、或替换新方案所带来的社会负面影响。

这里看一下蒙古文信息化产品的生态圈现状。



在能够自行安装第三方字体的 windows 和网页系统上，解决各种行业需求的蒙古文信息化产品还是很多的。也就说在你们不怎么关注到的领域，在电脑上用蒙古文工作、学习、生活是已经没有任何障碍。比如新闻、电视、出版、教育等各个领域。只要是解决方案里的所有产品都用了同一个生态圈里的产品，那么到目前为用的很好，不存在什么严重的兼容性问题发生。有点仅仅是跟其他文种一样错别字相同级别的错误。这些都通过辅助工具能够很好地解决。这个现象说明字体兼容性问题可观存在，我们也不是已经满足于现状，这里只是想表达现行编码方案可能不是大家想象的那样恶劣。

当然希望这次的会议能促进解决用户不能字形下载第三方字体的系统平台。如 iOS，Android。

5. 输入法

蒙古文是一种拼音文字，目前社会上存在两种输入法：全键盘输入法和整词输入法。无论哪种，用户在做输入动作之前，首先做的是通过大脑计算出想要输入法单词的字母结构！然后依据字母映射键盘来连续敲击键盘来输入法蒙古文。如果碰到一个按键序列对应多个单词时，需要使用自由变体选择控制符。也有能够省略敲击自由变体选择符的整词输入法。

基于上述蒙古文输入逻辑和过程，现行音码方案在输入法、再编辑等多个环节有先天优势。反观形码方案，虽然可以有类似于汉语五笔输入法，但是对于普通用户来说还是离不开整词输入法。

6. 排序

由于现行音码方案是基于蒙古文字母结构，所以在排序上几乎有天然的方便性。唯一特殊考虑的是单词中间存在自由变体选择符的情况。在这方面，形码方案需要更为复杂的计算，才能得到音码方案能够做到的效果。

这里再次说明，无论哪种方案，通过额外的辅助工具，都能弥补自己固有缺点，而达到另一种方案的先天优势。

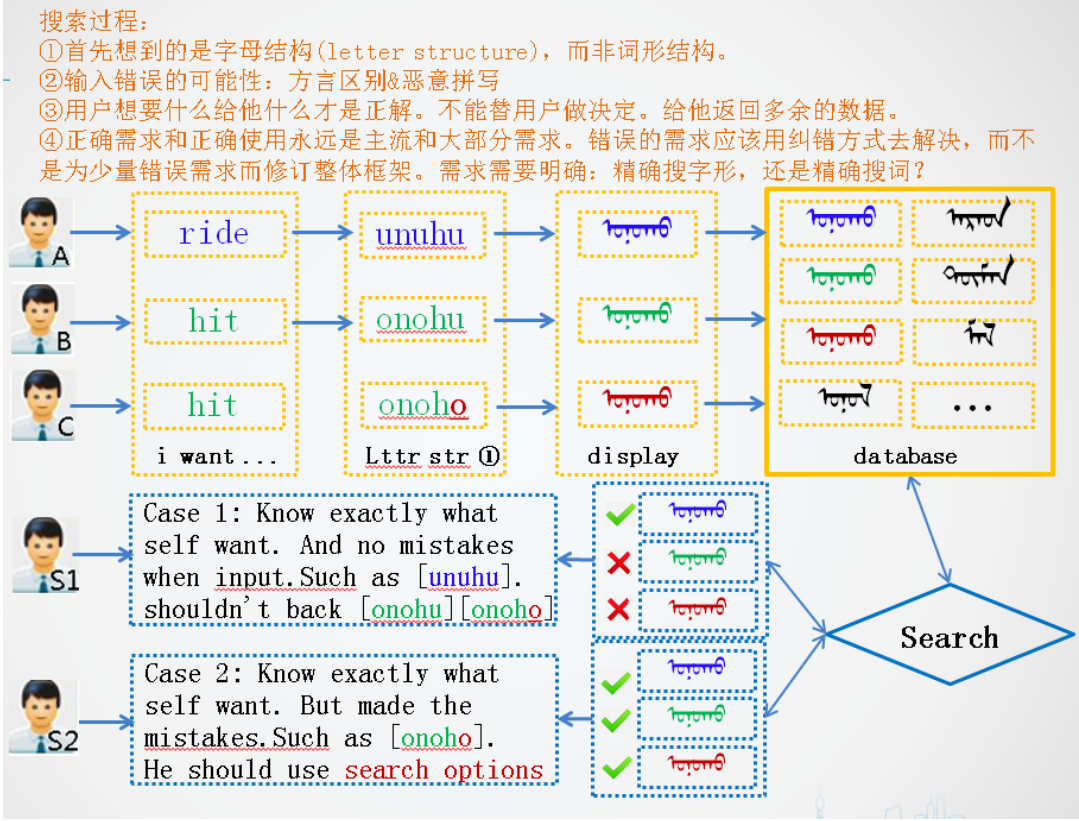
7. 搜索

就像任何一种方案都有优缺点一样，搜索只是现行音码方案的很小很小的问题点。这个跟形码方案为了追求词形无二义性而带来的排序复杂性一样的道理。

用户需求的出发点不同，功能需求也是有所不同的。如果只是关注单词字形，

那么形码方案是合适的。但是蒙古文实际使用者（尤其在自然语言处理上）从来不关注单词词形，而是关注点集中在字母结构上。蒙古语书面语（字母结构）和口语（发音）的差异是非常大的。所以对于蒙古语来说，单词的字母结构比什么都重要。在实际需求中，真正需要的是精确的字母结构搜索，而不是模糊字形搜索（模糊是指同形不同码）。

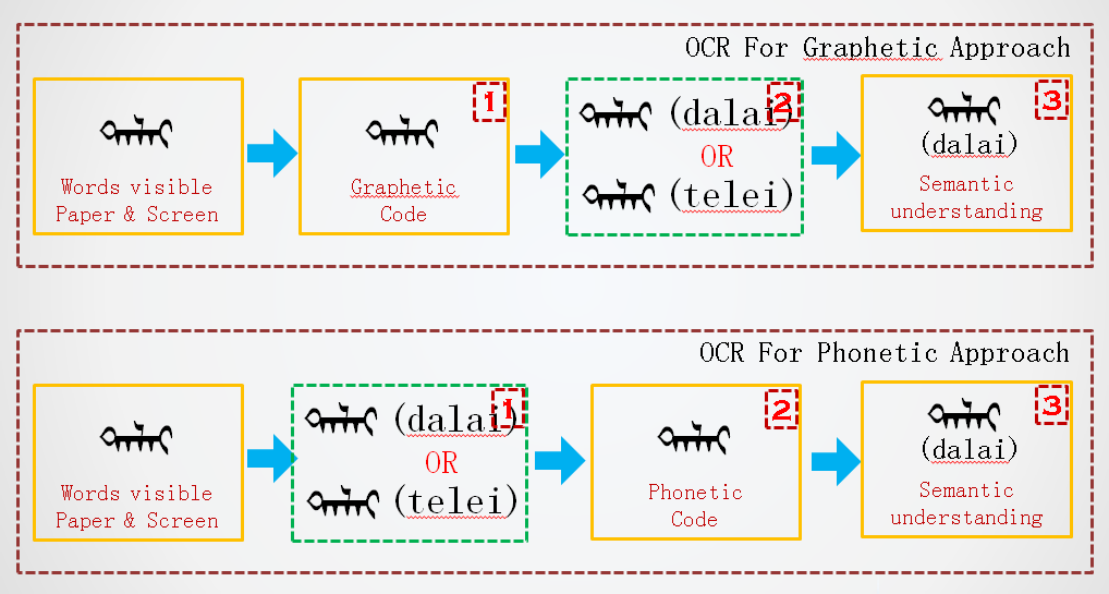
关于搜索问题上，相对于形码方案，现行的音码方案是把双刃剑。相对于形码方案不能精确搜索单词字母结构，音码方案可以很轻松地实现既能按字形模糊搜索（借助辅助工具），也能按字母结构精确搜索。



8. 词形到词码

这种应用场景的出发点是眼睛看得见的文字需求，最典型的案例是 OCR。对此如果应用的目的只是停留在讲眼睛看得见的词形识别转换成编码（这时是形码），那么形码方案相比于音码方案有绝对优势。

具体流程如下：



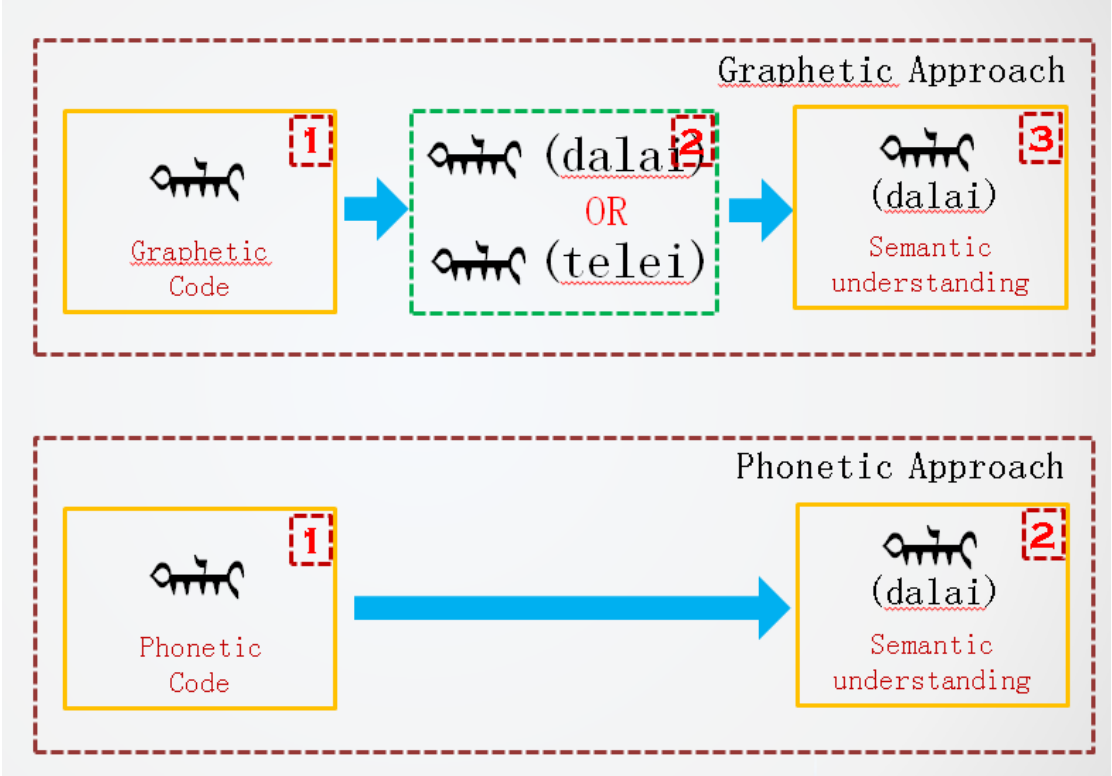
在形码方案下，从原始图片（眼见看得见的单词图形）到目标编码（形码）的过程到简单，也没有二义性。如同第一步，直接从图形一对一地映射出编码。但是计算机最终目的不是简单的计算出编码，还要进一步处理语义计算，如图第二步。这时就二义性的问题就会显现出来。为了精确计算语义，必须处理二义性。

在音码方案下，从原始图片（眼见看得见的单词图形）到目标编码（音码）的过程就比较复杂。直接通过二义性的复杂计算（如第一步），最后才打出实际编码（音码）。但是在后续计算语义时就没有了二义性，过程比较简单，如第三步。

从此，我们很容易对比出，在文字识别领域，形码并非比音码有优势。它只是在目标任务只是识别出编码层面具有优势，但目标任务为识别出语义的时候，它实际需要的计算量跟音码方案是相同的。

9. 词码到词义

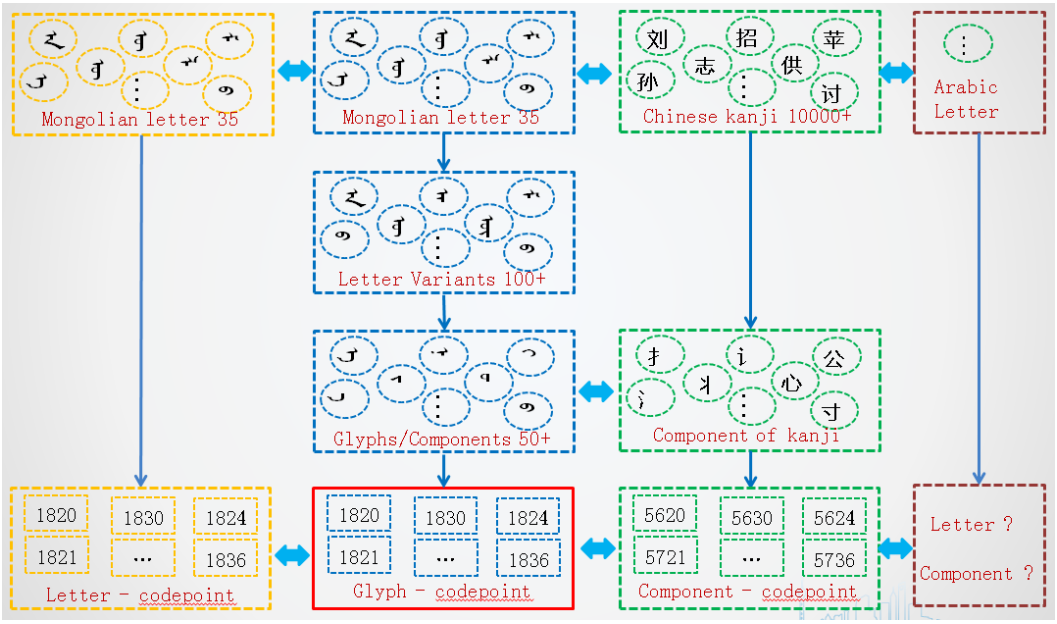
此类应用场景，其实就是从眼睛看不见的编码计算出语义的过程。此时音码方案的优势是比较明显的。因为音码本身已经携带了字母结构。从此图也很容易发现形码方案在减少同形异码的可能性的情况下，同时在增加同形异义的可能性。



10. 关于形码方案的个解读

形码方案倾向于注重形，减少形和码之间的二义性，但是在计算机处理时会增加形和字母结构之间的二义性。音码方案倾向于注重字母，减少码和字母结构之间的二义性，但是会增加码和形之间的二义性。

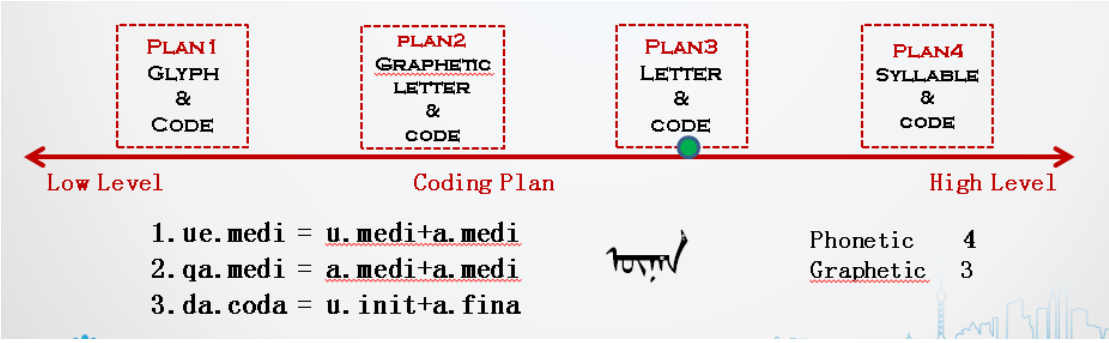
由此很容易发现，这种大家想极力回避的二义性不是来编码方案本身，而是这种文字本身的设计上。为了解决此问题，形码方案采取的是将蒙古文字母及其变体按照部件类型自由拆分形成了虚拟图形字母体系，并给虚拟字母体系编码。如同如下图：



关于形码，我们也应当考虑两个问题：

1. 形码方案真的解决了形和码之间的二义性吗？

由于在制作虚拟图形文字系统的过程中，还是保留了一部分蒙古文文字特征的部分字母，所以还是没有彻底解决二义性问题。所以它同样存在现行音码方案所具有的部分缺点。而二义性、安全性等等。



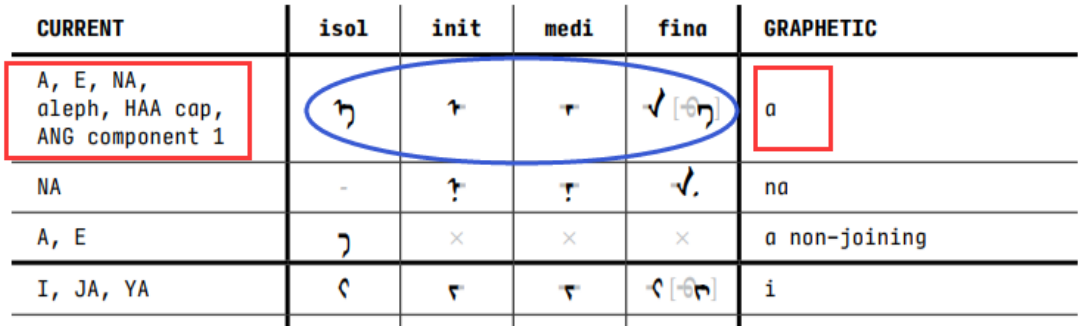
2. 为什么从感情上不容易接受形码方案？

阿拉伯文和蒙古文同样属于拼音文字。但是蒙古文本身的变形规则比阿拉伯文要复杂的多。如：

1. 蒙古文字母在词里一个位置上的变体不止一种。而阿拉伯文只有一种。

2. 在依据蒙古文正字法做蒙古文变体的自动选型时，除了依赖词里位置以外，还要依赖语法上下文，而阿拉伯文不需要考虑语法上下文环境。而目前，不使用 FVS 的阿拉伯文连写模型只适合于同一位置只有一种变体的文字系统。为了解决此问题，要么升级连写模型能够胜任更复杂语言，要么简化蒙古文变形规则。很

显然为了将蒙古文变体数量减少为一个，不得不将蒙古文字母变体以按字形部件来拆散打乱，重新归类制定出一套即不是字母、也不是字母变体的图形文字系统。如下页图。



CURRENT	isol	init	medi	fina	GRAPHETIC
A, E, NA, aleph, HAA cap, ANG component 1					a
NA	-				na
A, E		×	×	×	a non-joining
I, JA, YA					i

在上图中，一个图形字母可能是语言学的元音，也有可能是辅音。这种跨元音辅音的无差别编码方案使得图形字母系统已经失去了很多蒙古文的特征。包括元音、辅音等等蒙古文精华的部分。

在这种体系下很难实现阴阳性判断、拉丁转写、音节划分、元音辅音统计等。这些都是拼音文字特有的特点，也是它的优点。我们不能以方便性、易用性为目的，去剥夺语言文字特有的文化内涵。

这种不考虑语言文字内涵，而只是从字形变化上找规律的拼图游戏式方案很容易让人产生一种【不能因为一个技术问题而摧毁一个语言和文化。因为这种方式会把人都思维给改变。】的感觉。这是除了理性分析技术优劣以外的，一种纯粹感性的心灵上的抵触。

3. 大企业在等待稳定的编码方案问题

不应该认为只有简单易懂的形码方案是稳定的版本。规范而统一的音码方案也是一种稳定的版本。

（结束）