文字书写系统的计算语言学理论

导读

香港理工大学计算学系 陆勤

1. 学科背景介绍

当我们提到某个语言的【文字】一词时，普遍的理解不仅包括该文字所用的符号，还包括它的书写规律。在这里有必要解释一下文字和书写系统的不同。**文字**（script）在本书中特指某个语言的书写符号集，而**书写系统**（writing system）所指的不仅是作书写用的符号，还包括符号所用的**构件**（graphemes），构件的组成方法和相互之间的关系。举例来说，中文汉字是不同于英文的文字，其书写系统也有很大区别，一个是表意构件的两维排列，另一个则由字母拼写而成。日文作为一种书写系统，借用了其它语言的文字，但日文有其提独特的书写规则，因此其书写系统也是有别于它的书写系统。一般来说，如果文字符号所在选用的图形有随意性，该符号称为字母（alphabet letter），而字母没有拆解性，也没有构件的概念。但是，由字母拼写出来的文字不是随意的，某些固定拼法的**词位**（lexeme）带有语言学信息，才称为构件。在文字学中的**书写法则**（orthography）研究的是字或词的构件的书写规律[1]。本文研究的书写系统是关于词形所

[1]Orthography 在英汉字典里常翻成【正字法】，因为对中文来说，用【字】较为恰当。而对大部分的拼写文字来说， orthography 研究的构件以【词】和【词形】作为

能反应的语言信息。还要说明的是，本文所指的书写系统基本对象只是文字、文字的构件、以及从字或字母到词的构成。研究的对象并不包括语言学里的语法（syntax）和语义（semantics）。构件的最大集合只是到词，顶多会关联到复合词。

在文字学的范畴内，以往的研究较少使用计算语言学的各种方法和工具。本文的作者史伯乐（Richard Sproat）从事文语（text-to-speech， 文字到语音）转换的工作，是文语转换研究的拓荒人之一，参与了最成功的 AT&T 贝尔实验室的文语转换系统的开发。在处理单一语言的文语转换基础之上，为使该系统能够尽可能用最系统的方法扩展到其它文字的语音转换，就必须要寻找各种语言文字的共性与差异，并尽可能的用形式化的方法进行描述。只有这样，用计算机进行自动处理的程度才会提高。

本书中所阐述的理论就是在文语转换这个技术要求的大前提下产生的，其目的是要提出一种跨越不同文字而对书写系统进行形式化的描述方法(formal method)，或者说是要建立一个可操作的计算模型（computational model），用以表达从文字到语音转换所需要的特征和规律。为证明该方法能够用于不同文字，书中例举了超过十种以上的书写系统并在不同程度上描述了它们的共性与特点。 以中文作为母语的读者，并不需要对所提及的所有文字都有了解，有对英文的了解，就可以明白本书中所阐述的理论。实际上本书可以增进从事计算语言学研究和开发工作的读者对不同语言的文字的了解。也有助从事文字学和语言学工

对象，没有字的概念。为避免矛盾，在导读中在提到 orthography 时，回避直指【字】、【词】，而将 orthography 一词翻成【书写系统法则】。在用到【词形】和【构词】时，并不排除表意文字中【字形】与【构字】的书写规律 。

作的读者了解如何利用计算语言学的工具对所研究的对象进行形式化的描述。

内容提要

中文常用【阅读】一词来描述读书的过程，说明【阅】与【读】之间的紧密关系。我们要读出一篇中文文章时，在识别文字符号的基础上还需断词得当、语法语义理解正确，才能知道每个字词在文中的具体发音，这样才可能做到停顿有节，并合理的使用抑扬顿挫，使听众对文字所表达的意思能一【听】了然。从计算语言学或计算机技术的角度来看，我们可以把这个转换过程当成是词形作为符号到语音的映射。试想有多少语音信息是可以直接从文字符号中提取的，还有多少语音信息是文字本身无法界定而需要读者具有从其它途径获取的知识才能完成正确转换的？如果语音信息可以直接从文字中提取出来，也就是说文字到语音或者说词形到语音有直接的映射关系，在计算机系统中就可以用规则或赋值来实现。因此，文字中显含或隐含的语音信息及其规律性在很大程度上决定了文语转换系统所具有的知识如何获取以及转换的复杂程度。

本书以说明文语转换系统的可操作性问题为大前提，目的并不是要介绍不同的文字书写系统，最重要的理论论点都在第一章提出，其两个基本论点是：（一）词形到书写规则的映射存在**正则关系**（regular relation）；（二）一个特定语言的书写系统所表达的语言学信息具有**一致性**(consistency)。其它的章节主要是通过实例以不同的角度来对这两个论点作出详细的阐述和证明。第二章较详细的阐述了书写系统的正则性。第三章则详细说明了特定文字如何表达语言学信息以及所表达信息的一致性。第

四章介绍现代语言学中几种常用的文字体系分类方法，进而提出对文字书写系统的二维分类方法。第五章简要介绍如何用心理语言学的方法来分析母语读者进行文语转换的方式，并将本书所提出的理论与心理语言学的结论进行印证。第六章先讲解了文字与书写系统是如何被不同的文字借鉴以及承传的方式方法，然后给出不同文字中对缩写和数字的表述以及转换特性，最后对本书的内容做了一个总结。

2. 内容详细介绍

第一章 阅读器

本章一开始，以一个包括了数字、缩写、外来语（西班牙文的墨西哥词汇）的英文例句（*‘I need 2 oz. of Valrhona and 6 anchos for the mole’*），引出正确读音的困难，并且指出仅仅使用一般的语言学中词性和语法的标注信息来进行文语转换远远不够。很显然【阅】与【读】的过程中，形在前，而音在后，因此切入点是**书写系统法则**（orthography，简称**【书写法则】**）。对大部分文字来说，书写法则是指构件（grapheme）成词的书写规律，就是所用的字母（形的部分）在加上拼写的部分。不同语言的书写规则不同，从词形转换成语音的难度也不同。

本章首先引出心理语言学中常用的关于**书写系统深度**（orthographic depth）的概念，该概念用来表示从词形可以直接推断的语音信息的多寡程度。不同文字书写系统的深度不同，词形可推断的语音信息越多，直接转换越容易，则其书写系统的深度越浅。相反，词形本身可推断的语音信息越少，转换所需其

它信息就越多[2]，其书写系统的深度就越深。需要指出的是语言学范畴内所研究的信息不仅包括语音信息，文字的词形所包含的信息也不仅仅是语音信息[3]。换句话说，我们关心的是一个书写的文字包含多少语言学信息。从书写中得到越多的语言学信息，对文字要做的显性标注就越少。因此，本书提出**书写系统关联度**（orthographic relevance level）的概念。书写系统关联度说明的是一个文字的书写系统所能表达的语言学信息的程度。实际上每一个语言都有一个特定的书写系统联度，可以表示从一个文字的词形可以推断的语言学信息的多寡程度。

从文字的书写到其所具有的语言学中可以标识的属性关系（包括语音信息）可以用计算语言学中所谓的属性矩阵（attribute- value matrix）来表示。矩阵由各种语言学成分的信息组成，包括词形、语音、甚至语法语义[4]。词形中不同成分的语言信息的映射（mapping，也可称为对应关系）如果直接存在，该映射也可以在矩阵的架构内表示[5]。为使映射信息的表达更加直观，这些转换信息还可以用所谓的标示图(annotated graph)来表示. 标示图不仅可以表达属性矩阵的所有内容，还可以使不同属性之间的对应关系一目了然。在此基础上，本章定义了三种语言学信息之间的关系，包括一种时序（temporal）的重叠（overlap）关系、直接支配（immediate dominance）关系和

---

[2] 可以说成是计算机技术里称之为预处理（preprocessing）的工作越多。

[3] 比如英文的后缀就可以反映词性信息

[4] 这里说的是词形本身带有的语法语义信息，而不是篇章的语法语义信息。

[5] 举例来说，如果有些词的语法功能或语义不同时，其发音会有所不同，语法语义信息和语音信息就需要通过映射关系来表达。

直接路径优先（immediate path-precedence）关系.

文字的书写规则可以分成两个部分，一个是与词形本身有关的**图构规则**（graphic encoding rules），另一个是**拼写规则**（autonomous spelling rules）。一个重要的概念是所谓的**联接方式**（catenation）包括书写结构顺序以及对应语言学信息之间的关联。书写的结构顺序容易理解，有串联方式（concatenation）和二维平面（planar）方式两种。而语言学信息到词形的映射反映在这些信息是如何【串】在一起的（spellout）[6]。对应之前定义的三种语言学信息之间的关系，本章给出了三个相应的公理表达他们对词形的映射。而这些公理成立的前提都是基于词是由更小的语言构段组成的（segmental），比如词素、音素、在中文中则是部首、部件等。

在此基础上，本章提出全书的两个基本的理论论点：第一，书写系统关联程度到书写中间的映射存在**正则关系**（regular relation）；第二，一个特定语言的书写系统所表达的语言学信息具有**一致性**(consistency)。 正则关系是一种可以用数学来表示的简单映射关系，有了这种关系就可以用正则方式表达（regular expression），而可以用正则方式来表达的符号串都可以用计算机技术里的所谓有限状态转换器（finite-state transducer）来实现，因此第一点说明的是可操作性问题。正则关系有组合性，所以一连串的分部映射可以简化成一个单一的映射。正则关系还具有可逆性（invertible）， 也就是说从书写方法到书写系统关联程度也存在正则关系并且满足反向映射。有一

[6] 以中文【映】字为例，其部件的左右分配结构是二维平面式的，先写左部首【日】，后写右部件【央】。从语言学角度，左部首带表意成分，右部件带表音成分。

些非正则的限制性书写规则[7]，在计算机技术中可用有限状态接受器（finite-state acceptor）来实现。第二个点认为不同文字所带的语言学信息不同，但是，一个特定语言文字所带的信息具有稳定性，其程度不会随词汇的变化而改变。从表面上看，本书提出的理论对文字的书写系统是一种受限的理论，试想世界上有如此多的语言，但真正有与之对应的文字却不是很多。说到底，一个书写系统是给人用的，没有规律，就很难学会，所以受限是自然的。

本章最后引入全书所用的一些术语。

第二章 正则性

本章主要是通过介绍几种不同文字的书写系统来说明正则性不仅存在于基于字母的**拼写书写系统**（alphabetical orthography）而且也适用于其它的基于符号的**符号文字书写系统**（logographic orthography）。其中举例说明的文字包括中文汉字、韩字（Hangul）、印度的天城文（Devanagari）、还有杨松录苗文（Pahawh Hmong）。本章用了大量篇幅讨论中文汉字，因为中文作为典型的表意文字是一种典型的二维平面的书写系统。一个汉字由部件相对位置的结构安排组成，而部件可以带有语音、语义信息。书中对汉字进行拆解，并以形式化的方法进行正则描述。虽然在日常汉字教学和使用中很少有人直接使用这种数学的表示法，但在部件基础上描述汉字对写中文的人来说并不陌生[8]。

---

[7] 比如上下文限制就不能用正则关系来表示

[8] 实际上汉字结构符已有国际编码，表示汉字部件结构已相当方便。

韩字[9]（Hangul）作为一种表音特征的文字也是一种二维平面的书写系统。音素的书写顺序以及在二维平面的位置具有正则性。天城文（Devanagari）作为一种十三世纪才出现的文字源于婆罗米文（Brahmi），天城文是一种基于字母-音节（alphasyllabary）的文字，其字形到语音并不同构（non-isomorphic）。杨松录苗文的历史只有 50 年左右，使用语音字母书写而成。

本章给出了一个非正则文字的例子就是古埃及文。在古埃及文里，复数是通过重复相应概念字来表达的，而在书写时对概念字重复的次数并没有限制。由于随意的重复这种书写形式没有正则性，所以古埃及文是一个反例。不过公元前两千多年前，这种表示复数的重复写法已经被一个固定的重复符所取代，因此这种非正则的重复方式已不复存在。本章的最后特别提到篇章的书写顺序，比如中文传统上是从上至下，从右至左，因此书页的翻看顺序也是从右至左。实际上无论采用哪一种书写顺序，从正则表达上都不造成困难。

第三章 书写系统关联度与一致性

本章通过对几种不同语言书写规则的阐述，来说明不同语言书写系统关联深度的不同，包括词形变化的规律、所反映的语音及语法语义信息。文字书写中直接隐含的语言学信息越少或越模糊，其书写系统关联度越深[10]。但是，每一个特定语言的文字能够表达的这些语言学信息具有稳定的一致性。

---

[9] 也称为【谚文】

[10] 在计算机技术中可以想象成需要人为标注的信息较多。

第一节以阐述俄文和白俄罗斯文之间的比较和差异作为重点。这两种文字都用西里尔字母（Cyrillic alphabet）作书写符号，使用的书写特征也很类似，但有些规则却不尽相同。比如，两种文字的读音都有类似的元音轻音化（vowel reduction）读法，但词形的变化规律不同。在读俄文时，如果元音为非重音，其发音会发成轻音，但书写形式没有变化。白俄罗斯文在读音上也有类似的轻音化，但是轻音化在书写形式上能够直接反映出来。很显然这两个来自同一语系的文字，在元音轻音化方面其书写形式所带的语音信息不同，所以书写系统的深度不同，俄文较深，白俄罗斯文较浅。但是，俄文中在腭化音方面则用显性的书写变化来反映发音的变化（v 变 r）。每个语言的文字本身都有其特定规律[11]，因此有其固有的一致性。

英文作为世界上最普遍使用的文字，虽然被 Chomsky 和 Hale 称为最完美的拼写文字，但大多数人认为英文是一个书写系统深度较深而且拼写一致性较差的文字[12]。为了证明这一论点，本章使用了大量的篇幅来描述英文的一些拼写现象，所用的分析数据是从电子词典中选出了一千多有拉丁词根或希腊词根的英文词以及这些词根的各种变体，用来完成两项分析任务，第一是审视英文拼写的规律性，第二是测度需要多少人工词汇标注（lexical markings）才能够表示词形的书写特征（orthographic properties）。书中列出两种标注方法，其中一种标注被称为深层的词形标注，而另一个被称为浅层的。深层标注所需的词形到

[11] 读者如对某些语言不熟悉，不必过多注重其细节。

[12]实际上正因为英文是一个普遍使用的语言，从不同语言中直接引入的词汇也较多，因此词汇的不一致性较大并不足为奇。

语音的转换规则有 58 条，而浅层的有 69 条。不要以为转换规则多是好事，相反，因为多了规则，反而多了歧义性，因此需要更多的词汇层面的标注才可以消除歧义。值得一提的是英文的元音作为非重音也有轻音化现象，但没有词形变化，而浅层标注中进行的一个人工词汇标注恰好是用来标识轻音化了的英文元音。

本章对塞尔维亚克罗地亚语（Serbo-Croatian）中辅音清音化的特例（d 变 t，b 变 p）进行了分析并通过实验来寻找规律。值得一提的是其实验方法，第一步先准备不同的目标文本，在一定的控制速度下请人看读并录音。看读的人事先并不知道看读的目的，看读完成之后，通过语音处理的软件作为辅助工具进行自动分类，之后再作人工分析。结果证明，这种表面看似特例的语言现象实际上有规律可循，而且在词形变化上与其它情况并没有矛盾，因此从反例变成了正例。

本章还对双辅音字符串在拼写文字中对语音的影响进行了简要的说明，主要是基于 Nunn 对荷兰语中几个现象的分析，但是这种现象并不普遍，而且在音节上能够被切分，因此不会造成词形到语音转换的不一致性。

虽然拼写在很大程度上可以看成是语言学的某些构件到书写表示的转换，但还是会有一部分与语言学信息无关的词形本身要求的变化，这些变化规律可以称之为对词形外观的限制（surface orthographic constraints）。 本章以马拉加西文（Malagasy）为例说明了这种限制，而实际上这种限制同样可以用正则表达式来描述[13]。

---

[13]中文字的书写其实也有类似限制，举例来说，【土】字作为部件用在左边时（比如【堆】字）就要用提土，该变体对字本身的字义、字音都不产生影响，也不影响与

第四章 语言学成分

所谓语言学成分在这里指文字符号所带有的语言学信息，语言文字学对书写系统的研究最关心的其实就是这个问题。本章先重述了语言文字学中对书写系统的几个比较有影响力的分类，然后对中文和日文进行了一些系统性分析，最后也提及了几个稀有文字。

在语言学领域，Ignace J. Gelb 常被认为是文字分类的鼻祖。他认为使用构段的表音字母（segmental phonographic alphabet）在文字的发展史上具有革命性的意义，而文字的分类可以两点为轴，其中纯符号文字（logograph）[14]作为一点，表音的字母文字（alphabet）作为另一点，所有的文字分类都可以在这两点中找到其特定的位置。而 Jeoffrey Sampson 的重要贡献是认为文字作为符号可以传递两方面的信息，一方面是用来表达语言和语言形式的信息（glottographic），另一方面是直接表达概念本身（semasiographic）而没有语言学信息。实际上，到现在也没人发现任何只用直接表达概念的这种方式来书写的文字。中文在 Sampson 的分类中属于纯符号文字（logographic）。原则上说，文字的某些非表音的构件，不管代表的是完全的语素还是带有某些语义信息都可以称之为直接表达概念的构件，只是一般都会回避使用这个称呼而已。另外 Sampson 还是第一个将韩国的谚文称之为特征文字（featural alphabet）的人[15]。John DeFrancis 针对

其它字的关系， 因此，纯属字形外观的限制。

[14] 英文的 logographs 之所以称为纯符号因为没有可拆性，符号独立，也无需拼写。

[15] 这里指的特征其实是语音特征，这是因为韩字构件作为图形字母不是随意选的，

Sampson 的主张提出，一个完整的语言不可能仅仅使用直接表达概念的书写符号，因为没有语言学信息，一个文字的表现力有很大限制，充其量只能当作一组特定符号进行某种固定式的交流。DeFrancis 进而认为所有的文字都是表音的（phonographic）。他认为中文虽然表面上是表意的，但大多数中文字是形声字（morphosyllabic），所以中文也是表音的。DeFrancis 对文字的分类呈树形结构（arboreal，hierarchical），最上一级分为音节（syllabic）和构段（segmental）。 另外，DeFrancis 认为韩字不是特征文字，而是基于构段的（segmental）语音字母来拼写的文字，因为韩国儿童把一个韩字的成组音节作为一个完整的语言单元来记忆和使用的，他们并不知道语音特征与韩字的语音构段之间的关系[16]。实际上，韩字即使在表现语音特征时，也缺乏某些一致性。但是，不能否认韩文的书写系统是目前常用文字中设计最科学的。

本书的作者认为以树形结构来给文字分类有很明显的缺陷，因为很多文字多多少少都会含有一些语音信息和构段信息。因此，将文字在二维空间进行分类会更合理，其中的一维是语音的不同类型（type of phonography），另一维则表示其包含纯符号的程度(amount of logography)。当然以这种方式分类，维数有需要时还可以扩展，并不一定限制在二维之内。

本章的第二节用了很大篇幅来解释中文字的特征及分类，其中最有争议的是形声字，因为中文字的声部所表达的音缺乏一致性，有时甚至完全无用。即使 DeFrancis 也认为中文的表音部分

表音韩字构件的图形在某种意义上反映了发音的特征。

[16] 这个例子的证明属于心理语言学范畴，在第五章中有更多的论述。

不完美，但还是认为中文是主音的，原因之一就是中国人在见到一个不熟或不认识的字时，会很自然的以其认定的音部发音。即使是作为词素（morepheme）而不可单字使用的双字词,例如【槟榔】、【傍徨】，其字音也都由音部决定，而且双字的偏旁往往一样，标明其表意的功能。

日文的书写系统相对较复杂，日文往往被类归到符号文字（logographic）。日语最早借用汉字表达时，借用的方法有好几种，有一些仅取汉字之意而保留了日语的发音，有些则直接将汉字当一个符号来用而不理会汉字的字源，还有一些则是遵从汉字造字的基本原理创造出来的日本汉字，日文称之为国字（Kokuji），还有的汉字则直接用来当音字使用（宛字， ateji）。正是因为传统日文（以汉字为主）表音的能力较差，日文才会有表音的假名，并且在现代日文中，汉字的使用越来越少。

本章的最后一节介绍了几种以比较少见的书写方法表达语言学信息的文字。在叙利亚文（Syriac）中复数的表示是在所要拼写的词的一个字母的上方多加一点，但其作用则以词为单位，而且点的位置因应词中所包含的字母而定，复数并不改变词的读音，因此，这种词形变化的表达只会影响词意，而不影响发音。在早期的马来文（Malay）以及高棉文（Khmer）中用专门的后缀符号来表示对词根（base）的重复[17]。因为重复可作用于之前的一个词或一个词组，因此没有正则性，原则上无法用有限状态机来实现。但是，这种重复在实际应用时是受限的，也就是说作者不可能任意重复，在计算机处理时，可以通过词汇层面的标注进行学习，从而获取规律。另外有些文字中为表示某些字母不发

[17] 类似中文【高兴高兴】和【高高兴兴】，但重复由重复符来表示。

音，会在书写时加注免读符号（cancellation sign），免读符号实际上是一种字面上的语音显性标注。

第五章 心理语言学的成分

心理语言学研究的是人如何从文字中抽取语言信息以及如何从思维系统的语言转换成实际的文字。既然本书的目的是建立文字书写系统的计算模型并表述其与语言学的关系，理应衡量所提出模型的心理真实度（psychological reality），也就是要找出一般心理语言学的模型所具有的属性特征来验证本书提出的模型所具有的属性特征是否吻合。求证的方法可从两方面入手，第一个是**模型架构的一致性**（uniformity），第二方面是所谓的**双路径性**(dual routes)。在这里，双路径指的是读者在【阅】字时获取【读】音信息所用的方法（称之为路径）。心理语言学认为获取读音信息有两种主要路径，第一种通过词条对应（lexicon matching）完成，另一种则通过字形到音素（grapheme-to-phoneme）的**形音转换**完成。心理语言学关心的是哪种文字系统倾向于选用哪条路径。心理语言学中有一个关于**书写系统深度的假定**（orthographic depth hypothesis）。这一假定在最严格的形式下认为，应该至少存在一种书写系统深度够深的文字其读音只用词条对应的路径。与此对应，也应该有书写系统深度较浅的文字其读音只用形音转换的路径。作者认为，不同语言的书写系统深度会不同，字形到音素转换的规则性也会不同，但它们从词形到语言学对应的模式（manner of mapping）没有什么不同。至少从各种研究中可以得到两个共同的结论: 第一，对所有的语言来说，这两条路径都可供使用；第二，书写系统深度在最严格形式下的假定不成立，也就是说，可以证明所有的的书写系统确实都有使用双路径。

在心理语言学上路径使用的证明可以由两种实验来证明，一种称为**词汇判断**（lexical decision），另一种称为**命名**（naming）。给出一个测试语言的字符串，词汇判断实验要求母语读者判断该字符串是否是其语言中的一个词，而命名实验则要求读出该字符串的读音。对不同书写系统深度的语言进行实验的结果证实了母语读者两条路径都会使用。本章特别提到了为中文和日文进行的几个实验，用来验证这两个书写系统深度较深的语言在使用时，语音信息在这两种语言中也起着很大的作用[18]。

本章最后又介绍了一个类似神经网络的**联结网模型**（connectionist model）的心理语言学文语转换模型。这一学派的代表 Seidenberg 和 McClelland 所给出的所谓 Seidenberg - McClelland 模型（Seidenberg - McClelland Model）摒弃了双路径模型的两个基本模式，认为从思维开始，文语转换的过程包括了字形、语义、语音、上下文的信息，而这些信息的不同成分以一种类似神经网络的形式相互作用最后得出正确的发音。但作者对这个模型证明的认受性有怀疑，因为用以证明这一模型所作的实验有很大的局限性，所选择的数据限制较强（单音节），试验结果的准确性不够高，而且完全不能说明多音节中重音是如何处理的。

第六章 其它

本章简单的讨论了书中前几章没有讨论的四个问题。第一个部分陈述有关书写系统用于某个语言的适应性（adaptiveness）问

[18] 几个实验的具体设计过程并不复杂，在这里不作累述。

题。中文是世界上屈指可数的几个独立发展出来的文字(script)和书写系统(writing system)。其它文字大部分是承袭或借鉴了某个已有书写系统中的文字，经过对文字和书写规律的改造来适应其对应语言的要求。比如，希伯来文（Hebrew）和阿拉伯文都源于闪语（Semitic）语系，并承袭了闪语的基本书写系统。而像维文（Uyghur），则只是承袭了阿拉伯文字母的部分，犹太人的依地语（Yiddish）也是只承袭了希伯来字母，它们的书写方法因应自己的语言而发展出其独特的规则。一般来说承袭或借鉴一个文字书写系统有两种方法，一种是语音承袭（phonological faithfulness），另一种是词形承袭（morphological faithfulness）。书中例举了英国曼岛盖尔文（Manx Gaelic）如何承袭了英文的书写系统，并说明盖尔文对英文的承袭不仅是其表音部分，还有其词形特征。

本章第二部分讲解了荷兰文拼写规则的两次规范。书中主要讲述了如何规范荷兰文中复合词的复数形式。在荷兰文中，复合词的复数表示要求与复合词中第一个词的复数形式一致。1954年的第一次规范要求词形要与第一个词的可数性相关联，也就是与其词义相关联。因为某些词的可数性特征不清晰，这些复合词就很容易拼错[19]。而在 1995 年的第二次规范中则要求复合词的复数读音要与第一个词的复数读音一致而无需理会第一个词是否可数。这一要求表面上要求的是读音的一致，实际上却是要求复合词的词形与第一个词的词形一致，因此规范的是词形的承袭。

第三部分讲解了不同文字中数字所用的名称及其数字表达

[19] 较深的词形深度

方式的规律性。最早的阿拉伯数字的表示[20]是一种非自然语言（non-glottographic），是十进制的以位置（positional）来决定数字所代表的概念的书写系统。理论上说，在数字的描述中，有两个不同范畴的概念，一个是表达数字概念的名称或字符（name to represent a number or digit related concept），而表达数字概念的名称多寡与语言和思维习惯有直接关系[21]，另一个概念是对某个具体数字（specific number）的表达。中文对具体数字的描述以十进制为基础，书写顺序与阿拉伯数字书写的顺序一致，读写顺序与书写顺序也一致。古代玛雅人用的是二十进制，基本的概念符号数已经多了一倍。而马拉加西文（Malagasy）数字的书写与读出的顺序和阿拉伯数字表达的顺序整相反，也就是从最低位数读起。除了英文，其它日尔曼语系（Germanic）的数字在书写时会把个位和十位对调[22]。很显然，数字作为概念的名称是有限的，而每一个具体的数字对应的是一个独立的概念，不同的数字代表的是不同的概念，因此数字是无限的。之前曾提到无限的表达是非正则的，因此数字的表示是非正则的[23]。这里要强调的是，数字到其语言学的表示之间的映射是正则的。换句话说，已知一个数字（作为一个独立的概念），从它的数字表示到文字表述的对应关系是正则的[24]。

---

[20] 应该称为【印度-阿拉伯】数字（India-Arabic），因为现在使用的所谓【阿拉伯数字】起源于印度。

[21] 比如中文的【一】到【十】和【个】、【十】、【百】、【千】、【万】

[22] 英文数字个位和十位对调实际上只有十到十九。

[23] 这其实是间接证明了自然语言不可能是正则的。

[24]所以一个用文字表达的数字，可以用有限状态转换器（finite-state transducer）

第四部分描述缩写方式以及其语音转换。大部分的拼写文字都会用到缩写。但缩写有几种类型。第一种可以称之为**单词缩写**（abbreviation），主要是一个单独的词用缩略的方法拼出，比如 Blvd. 和 cm 分别是 Boulevard（大道）和 centimeter(公分)的缩写，而缩写的字母只是抽取原词的某些字母形成缩略词。还有一种缩略词和原词的字母无关，比如 lb 是 pound（磅）的缩写。单词缩写的特点是读音并无缩略，也就是说应该读原词的发音。另一种类型在词形上是取复合词的首字母进行缩略的，在英文常称为**首字母缩写**（acronym），比如 CIA（Central Intelligence Agency，美国中央情报局）, ACL (Association of Computational Linguistics，计算语言学会), NATO（The North Atlantic Treaty Organization, 北大西洋公约组织）等。在发音上首字母缩写又可以分成两种，其中 CIA 和 ACL 都采用字母直读，也就是读出每一个字母的名称，在本书里称为**字母直读**（letter sequence）。另一种是将缩写当作一个词而根据词的拼读规则来读音，比如 NATO。在本书里，可以称为**缩写拼读**[25]。实际上首字母缩写的读音对于一般的文语转换系统来说可看成是固定赋值，用属性矩阵标示即可。单词缩写的第二类也是固定的，只有单词缩写的第一类有任意性，因此需要用规则来【猜测】。

另外，本章还用了一小节讲解文字书写研究中的两种截然不同的观点，其中一种认为研究书写系统本身意义不大因为它充其量不过是模拟口语（spoken language），而另一种观点认为书写

找到它对应的数，相反亦然。

[25] 本书中，作者只将英文 acronym 用在缩写拼读上。

系统研究有意义，持这种观点的人又分为两派，一派认为研究的意义恰恰在于文字书写系统与语言的关系而另一派则认为文字是有别于口语的另一种沟通形式（separate communication form）。

作为对全书的总结，作者在最后强调本书提出了一个书写系统法则（orthography）的理论，阐述了文字书写中标注语言学信息的方法，其能够表现的语言学深度，以及语言学信息与词形之间映射的约束（constraints）。以往的研究注重对特定文字的约束，作者的工作以及所提出的理论则是对目前研究语言学信息与文字书写词形关系中最具系统性的尝试，并希望此项工作能够起到抛砖引玉的作用。

3. 评论及对读者的建议

汉字作为世界上最古老的文字之一，在历史的长河中不断的演变发展，并为其它的语言文字以不同的形式承袭和借鉴，证明了汉字有极强的生命力。汉字及其演变发展的研究在使用中文的国家和地区源远流长，在学习中文的热潮席卷全球的同时，对中文汉字的研究也远不限于使用中文的国家和地区。但从另一个角度来看，目前国内在文字学方面的研究重点比较局限于中文本身，对于文字和书写系统总体的研究较少， 介绍这方面工作的书籍更是可怜，因此，很高兴本书能够在中国发行，对有兴趣的读者提供参考。

在今天的全球化的大前提下，特别是互联网时代技术的发展，我们有更多的机会接触到不同的文字，因此有必要对于文字书写系统有一个较全面的了解。对一般的文字工作者和对文字发展演变有兴趣的读者来说，阅读本书可增加我们对不同文字书写系统

的了解，因此读者不必拘泥对某个特定文字细节的了解。而对于从事计算语言学研究和自然语言处理开发的工作者来说，这本书不仅仅可以起到帮助我们提高专业知识水平和视角的作用，还可以增进我们对研究工作和系统设计的总体考虑，特别是对其他语言的可宽展性有所启发和提高。

值得指出的是，本书作者的工作主要是在计算机领域，因此，书中用形式化方法（formal method）进行的描述较多，希望读者不要【望】而却步，有些看起来枯燥的部分可以略过。其实本书对不同文字系统的描述简明扼要，不失为一本简明的参考书。延伸阅读指南部分还会列出几部比较详细的以文字介绍为主的参考读物。另外，本书带出了在语言学中研究文字的两个主要方向，一个强调文字本身的特点和特征以及文字所含的语言学信息，而另一个强调的是读者如何学习、认知和使用该文字。这两个方向对有志从事语言学、心理学、计算语言学、和自然语言相关的计算机应用的人来说，具有一定的学科方向的指示。

最后还要指出，本书研究的对象在拼音文字里是在词的范围之内的，对表意文字，则基本是在字的范围之内的。但是，对文语转换系统来说，这个范围是必要而不充分的。显而易见，在文字从【阅】到【读】的过程中所要求的抑扬顿挫（prosody，音调）是不可能在词的基础上解决的。英文词 produce[26]是一个重音不同而意思和词类都不同的例子，它的正确读音原则上可以从其语法功能上进行判断。再举一个简单的中文例子，当你见到汉字【乐】时，如果没有上下文信息，你是无法判断它的正确读音

[26] produce 可为动词，意【产生】， 重音在 u 上；也可为名词，意【农产品】或【蔬菜】，重音在 o 上。

的，这些都说明了一个好的文语转换系统不能只用词形和其对应的语言学信息。香港有一位有名的极地探险家名叫【李乐诗】，如果没人告诉你怎样读她的名字，大概谁也不知道自己的【猜】读是否正确。在这里，【乐】字与上下文的【诗】搭配时，两个发音都有意义。这个例子实际上带出一个文字表达的局限性问题，当文字用来表述口语时，口语中的一些信息是不能完全还原的。因此，本书提出的理论不仅是一个对文字书写系统的受限的（constraint）理论，而且也可以说是一个有限（limited）的理论。

4. 延伸阅读指南

4.1 英文延伸阅读书籍

4.1.1 A Study of Writing, by I. J. Gelb, The University of Chicago Press, 1963

这是最早的一部对文字书写进行系统性阐述的书。本书对世界文明史中文字起源、演变作了比较全面的介绍，比较强调文字的历史、演变以及语音信息对文字演变的影响，是本很好的入门和文字历史的普及读物。

4.1.2 Writing Systems, by Geoffrey Sampson, Hutchinson & Co. Publishers Limited, 1985

这是继 I. J. Gelb 1963 年出版的关于书写系统的书之后的又一本对文字书写系统的专著。作者对汉字、韩文、以及日文的造诣较深，因此从主观上没有对字母文字的倾向性。作者在本书里提出韩字以基于语音特征的文字。对文字学的全面发展有很大贡献。

4.1.3 Visible Speech, by John DeFrancis, University of Hawaii Press, 1989

本书作者对文字书写系统的分类比较有兴趣，因此在介绍不同文字体系的基础上也介绍了几种分类体系，并从文字作为沟通工具的角度进而阐述了文字系统的共性。

4.1.4 The Writing Systems of the World, by Florian Coulmas, Basil Blackwell, 1989

这本书重点介绍了几个重要的文字体系，从语言学角度讨论不同文字体系的特点、特征、以及语言和文字之间的差异。是一本不错的文字学基础读物。

4.1.5 Writing Systems – An introduction to their linguistic analysis, by Florian Coulmas, Cambridge University Press, 2003

这本书顾名思义，从语言学角度阐述文字书写系统的语言学信息，特征，分析，并提到了心理语言学和社会语言学的领域对文字书写系统研究的问题，较适合从事语言学作和计算语言学研究和工作的读者。

4.2 其它英文相关书籍

4.2. 1 Writing Systems - a Linguistic Approach, by Henry Rogers, Blackwell Publishing, 2005

这是一本不错的介绍不同文字系统特点的教科书。

4.2.2 Information, Knowledge, Text, by Julian Warner, Scarecrow Press, Inc., 2001

这是一本简要的介绍文字、文本与形式化关系的书。

4.2.3 Written Language and Psychological Development，by Leonard F. M. Scinto, Academci Press, Inc., 1986

这是一本心理语言学相关的书籍，有对文字的心理发展模型（development model）方面的阐述。

4.2.4 Speech and Language Processing, an Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, by Daniel Jurafsky and James H. Martin, Prentice Hall, 2000

有关计算机技术中的正则表示以及有限状态转换器（finite-state transducer）的技术书籍，多不胜数。但这部书集中讲解了和自然语言处理相关的计算机技术，包括了本书所需的描述与实现所需的工具和技术。

4.3 中文相关书籍与文献

4.3.1 谢清俊 —【汉字的字形与编码：从缺字问题谈汉字交换码的重新设计 — 第一部分】，汉字字码与资料库国际研讨会，京都•东京， 1996 年 10 月 4 日，http://cdp.sinica.edu.tw/paper/1996/19961004_1.htm

4.3.2 饶宗颐 —【符号初文与字母：汉字书】，香港商务印书馆，1998

4.3.3 庄德明、谢清俊 —【汉字构形资料库的建置与应用】，汉字与全球化国际学术研讨会，台北，2005 年 1 月 28-30 日，http://cdp.sinica.edu.tw/cdphanzi/documents/T9401.pdf

4.3.4 黄居仁 —【汉字知识表达的几个层面：字，词，与词义关系概论】，汉字与全球化国际学术研讨会，台北，2005 年 1 月 28-30 日，cwn.ling.sinica.edu.tw/churen/漢字知識表達的幾個層面.pdf

4.3.5 周亚民，黄居仁 —【字义符知识结构的建立】，第六届词汇语义学研讨会， 厦门，2005 年 4 月 21-24 日，cwn.ling.sinica.edu.tw/churen/漢字知識表達的幾個層面.pdf

4.4 其它相关英文文献

4.4.1 Chou Ya-Min and Huang Chu-Ren, “Hantology: an Ontology based on Conventionalized conceptualization, proceedings of Ontolex 2005, October 15, 2005, Jeju Island, South Korea, http://www.aclweb.org/anthology-new/I/I05/I05-7002.pdf

4.4.2 Lu Qin, "The Ideographic Composition Scheme and Its Applications in Chinese Text Processing", *18th International Unicode Conference*, April 24-27, 2001, Hong Kong, pp(B12)1-17, http://www.comp.polyu.edu.hk/~csluqin/paper01Unicode20Paper1.pdf

4.4.3 Lu Qin, "Ideograph Variants - What they Are and How to Handle Them". Proceedings of 21st International Unicode Conference, Vol. 2, Dublin, Ireland, 14 - 17 May, 2002, Unicode Consortium, pp.C1701-C1709 (2002), http://www4.comp.polyu.edu.hk/~csluqin/paper02Unicode21Paper2.pdf

4.4.4 Lu Qin, Xing Hongbing, Li Yin, Li Ngai Ling, Chan Shiu Tong, "The Hong Kong Glyph Specifications for ISO 10646's Ideographic Characters". Proceedings of 21st International Unicode Conference, Vol. 2, Dublin, Ireland, 14-17 May, 2002, Unicode Consortium, pp.C1601-C1617 (2002), http://www4.comp.polyu.edu.hk/~csluqin/paper02Unicode21Paper1.pdf

4.5 其它相关资源

4.5.1 汉字结构信息查询系统：
http://glyph.iso10646hk.net/ccs/ccs.jsp?lang=zh_TW

4.5.2 中文异体字查询系统：http://dict.variants.moe.edu.tw/