汉语方言学的计算机应用

(Application of Computer in Chinese Dialectology)

复旦大学中文系　沈榕秋

内容提要：

本文根据作者自己研究经验，说明了计算机在汉语方言学中的重要作用。介绍了汉语方言的计算机语料库、汉语方言的计算机定性描写，叙述了汉语方言的计算机定量研究的基本步骤、研究范围。

Abstract

Based upon the author's own experience in research, the article reveals the importance of computer in its application in Chinese Dialectory. It is an introduction to computerised material data base and qualitative analysis of Chinese dialicts. It provides description of the major proceedings and research scopes in the computerised quantitative analysis in Chinese dialects.

语言学对于计算机科学起着至关重要的作用，这是众所周知的。相反，计算机科学对于语言学研究也能起着重要的作用。计算机科学的需求刺激着语言学的发展，这是间接的作用。而语言学者将计算机应用于具体的语言研究，从而革新了语言学的研究手段，拓宽了语言学研究的领域，解决了过去难以解决的问题，这是直接的作用。语言学用于计算机科学研究，计算机科学用于语言学研究，有人将此二者统称为计算语言学。因此可以说，计算语言学是计算机科学和语言学共同发展的纽带。　　在传统的方言学研究中，从语料的收集、整理到得出音系表、同音字表、词汇表等，整个过程都是手工操作的，这需要花费大量的劳动力。而且研究范围也受到很大的局限，它只能停留在定性研究上，即分析方言的成分，而不能进行定量研究，即分析方言成分的量。方言定量研究在语料的处理和计算时的工作量更是大大超过定性研究，以至人力难以胜任。

如今有了计算机，它不仅可以支持我们进行方言定性研究，大大减轻我们的工作量，提高研究结果的准确性；而且使我们终于能够从事方言的定量研究。下面我们将要具体地谈谈汉语方言的计算机研究。

一、方言语料库的建立

建立方言语料库是进行汉语方言计算机研究的前提。所谓建立方言语料库，简单的说，是将方言材料按照一定的编排格式输入计算机，使这些材料在计算机中形成一个有序的集合。有了语料库，研究者才能够通过计算机对材料进行查询、再编排、互换、计算和打印等操作。

建立方言语料库可以在ＣＣ－ＤＢＡＳＥⅢ系统上进行，这是一种计算机汉字数据库管理系统。在这个系统上，方言材料的输入和运算都比较方便。上海师大潘悟云先生用这个系统开发了“ＴＤＰ系统”，此系统可用于方言语音材料的

快速调查和输入。一种方言语音材料从调查到输入用一、两天时间就可以完成。

在早期，由于计算机还不能输入国际音标和一些特殊的方言字，甚至不能输入汉字，所以方言材料是以代码的形式输入计算机的。９０年我们在ＣＣ－ＤＯＳ的基础上开发了“ＦＤ汉字库”，使计算机不仅可以输入一般的汉字，而且可以输入国际音标和一些特殊的方言字，从而使建立的方言语料库既直观又易于进行计算机操作。

９０年，为了研究上海现代方音的变化，我们在上海市南市区对老年组、中年组、青年组、少年组四个年龄组进行了方音抽样调查，每组抽查五个人，共计二十人。调查的内容是上海话中常用字的读音，共计3477字。我们将收集来的20个人的语音资料分别输入了计算机，录成20个原始语料库。库中每个字为一条记录。下面就是一个人的语料库的一部分内容：

记录号	字	声母	介音	韵腹	韵尾	声调
320	驴	l		u		23
321	吕	l		i		23
325	侣	l		i		23
323	旅	l		i		23
324	虑	l		y		23
715	悔	h	u	e		34
722	会	k	u	e		34

语料库建立后，为了各种目的，要在计算机上将这些原始的语料库作进一步整理，生成若干新的语料库。在对上海现代方音变化的研究过程中，我们将20个原始语料库整理成了声母、介音、韵腹、韵尾、声调等语料库。如声母语料库（部分）：

记录号	字	L1	L2	L3	L4	L5	Z1	Z2	Z3	Z4	Z5	Q1	Q2	Q3	Q4	Q5	S1	S2	S3	S4	S5
9	大	d	d	d	d	d	d	d	d	d	d	d	d	d	d	d	d	d	d	d	d
10	罗	l	l	l	l	l	l	l	l	l	l	l	l	l	l	l	l	l	l	l	l
448	荒	f	f	f	f	f	f	h	h	h	h	h	h	h	h	h	h	h	h	h	h
449	慌	f	f	f	f	f	f	h	h	h	h	h	h	h	h	h	h	h	h	h	h

这是声母语料库的一部分。L、Z、Q、S分别代表老年、中年、青年、少年，L1表示老年组的第一个人，其余类推。这些语料库分别用于对声母、介音、韵腹、韵尾、声调等的运算。

我们现在已用类似的方法建立了吴方言的三十多个方言点的语音语料库和词汇语料库等。

二、汉语方言的计算机定性描写

过去汉语方言定性描写的主要工作是列出方言的声母表、韵母表、声调表、声韵调关系表、同音字表等，比较不同方言之间的或同一种方言各个时期的声韵调变化情况等。这些工作不仅相当烦琐和耗费时间，而且在处理过程中非常容易出错，令人望而生畏。现在这些工作都可以由计算机去完成。我们可以设计程序，让计算机对语料库进行处理，在很短的时间里生成我们所需要的上述各类表格，并将这些表格打印出来。计算机在处理过程中是不会出错的。因此，这既能大大减少方言学者的工作量和时间，又能提高方言定性描写的准确性。

我们在研究上海现代方音的变化时，曾与宁波、绍兴、苏州、扬州、盐城方言以及普通话等语音进行了比较，因为这这些话对上海现代方音的变化有不同程度的影响。下面所列的就是我们让计算机生成的上海现代方音变化与这些话的声母比较表（部分）：

代表字	波	颇	婆	觅	火	呼	肤	否	符
老年组	? b / p	p	b	m	Φ	Φ	Φ	Φ	β
中年组	p	p	b	m	Φ / h	Φ / h / f	Φ / f	Φ	β / v
青年组	p	p	b	m	Φ / h	Φ / h / f	f	f	v
少年组	p	p	b	m	h	h / f	f	f	v
宁　波	p	p	b	m	h	Φ	f	Φ	v
绍　兴	p	p	b	m	Φ	Φ	f	f	v
苏　州	p	p	b	m	h	h	f	f	v
扬　州	p	p	p	m	x	x	f	f	f
盐　城	p	p	p	m	x	x	f	f	f
普通话	p	p	p	m	x	x	f	f	f

从这个表中我们可以一目了然地看出上海方音的各个音位变化及其与有关方言的一致化过程。如：从老年到少年，“肤”的声母经历了由〔Φ〕变为〔f〕的过程，在中年组中经历了既可读〔Φ〕，也可读〔f〕这样一个过渡阶段。表中所有其他方言以及普通话都读〔f〕，可见上海方音中“肤”声母的变化是受这些方言的影响的结果。有了这些变化对照表，我们就可以比较清楚地看到上海现代方音的声韵调变化的全貌、全过程，以及各方言和普通话对它的影响。

三、汉语方言的计算机定量研究及其步骤

汉语方言的定量研究，也叫计量研究，它是根据统计学的原理对汉语方言成分所进行的系统的量的分析。这是随着的计算机发展而在汉语方言学中的产生的一种比较新的研究方法。

不同方言之间的差异以及同一种方言内部的变化等，不仅表现于系统成分（如声、韵、调）及其组合关系的差异与变化上，而且也表现于这些成分在语言中所占的量的差异与变化上。譬如：上海话中的韵母〔iE〕变为〔i〕，常用字只有一个“念”字；〔i〕变为〔y〕，常用字却有十多个，如“徐”“序”；有的变化甚至有上百个常用字。因此，要全面反映方言发展变化，既要研究方言的成分，也要研究成分的量。而一种方言的音系往往有上百个成分，这些成分之后有几千个常用字，当对多种方言的或同一种方言不同时期的常用字逐一比较，分析其成分的异同，并且进行计算时，其工作量以及所要花的时间是难以想象的。所以传统的方言学中只有人进行方言成分的分析，而没有人进行成分的量的分析。可是，有许多问题又非得通过定量研究去解决。

计算机的出现，以它超凡的容量、记忆力和运算速度，才使我们有可能进行方言的定量研究。现在国内外已有一些学者正在从事汉语方言的计算机定量研究。但真正了解它的人还很少，而且其方法本身也未臻成熟。下面根据我们的研究经

验介绍一下其主要的步骤。

1、材料的收集

汉语方言定量研究可以是通过直接调查的方式，也可以间接地利用别人的调查结果去收集材料，但有一个基本原则，这就是要求所收集的材料尽可能准确地体现总体的面貌。可是作为一种方言总体的说话人往往有成千上万，我们不可能去收集所有这些人的语言材料。为解决这个矛盾，必须采取抽样方法，即根据统计学的原理从方言中抽出若干个样本，以代表整个方言。一般来说，抽样越多，误差就越小。为了减少误差，往往要同时分析每个方言中的几个说话人的材料。

2、建立方言语料库

比较的方言越多，样本也就越多。样本越多，人工也就越难进行处理。所以，当材料收集好了，接着就要将这些材料输入计算机，建立方言语料库，以便让计算机去进行数量分析。建立数据库实际上也是对方言材料的量化，即使材料转化为适于计算机操作和数量分析的形式。

3、算前加工

为了提高数量分析的准确度，常常在计算前要进行材料算前加工。譬如，有的要增加必要的模糊度。定量研究对材料的一致性要求比较高，相同值就要用相同的标记。如果用不同的标记，就会被当作不同的值来计算。当材料的研究范围比较小时，最好是由一个人去收集这些材料，这样易于避免记录标记不一致的现象。但材料往往是通过两个以上的人去收集的，有时还要用第二手材料。这些材料放在一起时，就可能出现用不同的音标记录相同音值的现象，这将影响计算结果的准确度。为了尽可能避免这种情况，我们可以对材料中的个别标记作必要的调整，如［io］、［yo］一律记作［yo］；声调上，可以将五度记音化成三度记音，等等。这样增加了标记的模糊度，却有利于提高定量研究的精确度。

4、数量分析

在算前加工完成后，就可以进行方言的数量分析了。定量研究中的数量分析方法有许多种，目前用于方言定量研究的主要有相关系数分析，聚类分析，主分量分析等。相关系数分析是对方言各成分进行量的比较和计算，得出相互之间的相似或相异程度值。譬如：我们曾经在研究中得出，上海方言的老年组与宁波方言声母相关系数是0.7983，即有79.83%是相同的；韵母的相关系数是0.5389；声调的相关系数是0.5798。聚类分析是根据方言的相似和相异程度，即相关系数的值，将多个方言分成若干个类。主成分分析也是一种描写方言相似或相异程度关系的方法。它在相关系数分析的基础上，重新组织数据，使原来变量的高维空间变成低维空间，而有关信息损失尽可能得小。分析结果是图形，方言之间的相似或相异关系反映在图形上，比较相似的方言聚在一起，比较直观。此外，其他一些数量分析方法也可以根据具体的需要用于方言的定量研究。

每种数量分析方法往往有多种数学公式，不同的公式有不同的作用。可根据具体情况选择公式，以提高数量分析的精确度。

数量分析的计算可以通过现成的计算机软件来完成。这类软件比较多，其中有SPSS、SAS、BMDP等。有些数量分析，一般的研究者完全可以根据数学公式自己设计程序去完成，如相关系数分析等。

5、得出结论

数量分析的结果是由计算机给出数字和图形，我们可以根据这些数字和图形做出结论。但为了确保结论的正确性，需要佐以别的证据。

四、汉语方言计算机定量研究的范围

汉语方言计算机定量研究的对象可以是方言语音、词汇、语法以及风格等。但现在只是对方言语音、词汇进行研究。相信不久以后一定会有人对方言语法、风格等进行计算机定量研究。目前在语法上面临的是用哪些语法现象去体现方言的整体语法，这个问题的解决可能有待于语法定性研究的进一步深入。

汉语方言计算机定量研究既可以对方言作共时的研究，也可以作历时的研究。

从方言的共时角度看，方言之间存在着相似程度的差异。北京人一般听得懂西安方言，却听不懂福州方言。这是因为北京方言与西安方言相似程度高，而与福州方言相似程度底。由于方言之间的相似程度不一样，方言存在着亲疏关系。当比较多种方言的亲疏关系时，这种关系就会显得很复杂。计算机定量研究可以通过求出语音、词汇或语法的相关系数，以相关系数表的形式全面描写多种方言之间的亲疏关系，或以聚类树形图、主分量分析图的形式简单直观地描写这种关系。

从历时角度看，方言与生物现象一样，相互间有遗传上的联系，这种联系就是亲缘关系。汉语各方言是从一个母语不断分化出来的，有的相互之间分化得早，有的分化得晚。早分化的亲缘关系远，晚分化的亲缘关系近。方言的亲缘关系可以用计算机定量研究的方法推求。在这方面王士元、沈钟伟先生曾参照Fitch的运算方法进行了尝试。他们将方言间的词汇相关系数转换成距离，然后根据距离用数学公式找出树状结构和树的根，从而推求出方言的亲缘关系。(王士元、沈钟伟，1992)

汉语方言的分区问题一直是有争论的。我们认为方言分区可以分为两种，一是历时性的分区，一是共时性的分区。历时性的分区是根据亲缘关系的远近进行分区，这种分区图可以反映出各方言之间的历史关系。共时性的分区是根据亲疏关系进行分区，而不考虑各方言之间历史上的联系，这种分区图可以反映出各方言之间的相似程度，即可懂度。过去分区方法有许多种，但往往是历时与共时不分，所以现有的方言分区图上既看不出各方言之间的历史关系，也看不出听懂度的关系。现在所分的十大方言区，北方方言区内部相互之间都可以听懂。如果外行人根据这一点判定其他方言区情况也是如此，那就错了。实际上其他方言区内部往往有相互之间听不懂的，甚至次方言区的内部也有相互听不懂的。因此，我们有必要研制出一部共时性方言分区图，以反映各方言之间的听懂度。计算机方言定量研究既可以研究亲缘关系，也可以研究亲疏关系，所以，历时性和共时性的分区都可以用计算机定量研究的方法进行研究。亲缘关系的计算机定量研究比较复杂，还有待进一步的探索。目前，通过亲疏关系的计算机定量研究进行共时性的分区，是完全可行的。我们可以通过实验的方法，找出一个方言相似系数值，作为听懂与听不懂的界线，然后根据这个相似值对方言分区。也可按照基本听不懂——大部分听不懂——小部分听不懂——基本听懂这四个层次，确定相应的相关系数值，将方言分成大方言区、次大方区、次小方言区、小方言区。用计算机定量研究的方法进行方言共时性分区的优点是客观，符合人的语感；对方言分区者来说，也不至于象以往那样陷入那种“剪不断理还乱”的境地。

方言的划界对方言学者来说也一直是个棘手的问题。方言计算机定量研究也可以在这方面进行尝试。我们认为下面的方法是可行的，即在相邻的两个（或者三个）方言区中，根据相关系数各找出一个代表性的方言点，假设找出的分别是

A、B两个方言点，然后比较方言区之间其他方言点与这代表性方言点的相似系数。假如，某方言点与A的相似系数值最高，这个方言点就归入A所代表的方言区；另一方言点与B的相似系数值最高，这个方言点就归入B所代表的方言区。当所有点都有了归属，方言区界也就水落石出了。比较的点越密，区界也就越精确。

方言计算机定量研究还可以用于研究方言的变化速度。我们曾对三十年代到现在的上海方音变化速度作了计算，结果是：声母变化了11.64%，韵母变化了16.99%，声调变化了2.40%，整个语音则变化了10.04%。而同一时间里的绍兴方音的变化是：声母变化了2.47%，韵母变化了2.95%，声调基本没变。（沈榕秋、陶芸，1992）

此外，计算机定量研究还可以是用于方言的分化年代、相互影响程度（沈榕秋，1993（A））、解释某种语言理论（沈钟伟，1990）、考证古语（见《计算语言学导论》上编第五章"历史音韵学研究"，陆致极）等方面。

以上所谈的计算机定量研究的范围，全都超出了传统方言学的研究范围。我们相信随着研究的深入，这个范围势必会得到进一步的扩大。

可以肯定，随着时间的推移，从事汉语方言的计算机定性和定量研究的人会越来越多，方言学必将会有很大的收获。计算机科学正在使汉语方言学发生着深刻的变革。

参考文献

郑锦全，〈汉语方言亲疏关系的计量研究〉，《中国语文》88年2期

马希文，〈较方言学中的计量方法〉，《中国语文》89年第5期

陆致极，〈汉语方言定量分析的理论模型〉，《现代汉语定量分析》陈原主编，上海教育出版社，1989年

沈钟伟〈词汇扩散:人口调查和数学模式〉，JOURNAL OF CHINESE LINGUISTICS, Vol.18, No1.1990年

王士元、沈钟伟，〈方言关系的计量表述〉王士元、沈钟伟，《中国语文》92年第1期。

沈榕秋、陶芸，〈上海现代方音的变化速度〉，《复旦学报》92年第4期

沈榕秋，〈上海现代方音变化与移民方言、权威话关系的计量研究〉，将发表在JOURNAL OF CHINESE LINGUISTICS, U.S.A. 1993年a

沈榕秋，〈汉语方言的计量研究〉，《语文研究》93年第4期b

陆致极，《计算语言学导论》，上海教育出版社，1990年。

童忠勇，《统计分析软件SPSS/PC+》，陕西人民教育出版社，1990年。