语言科学 2008 年 3 月
ISSN 1671 - 9484 CN 32 - 1687/ G 第 7 卷第 2 期(总第 33 期) 113 - 126

汉语书面语体庄雅度的自动测量 *

冯胜利[1,2] 王 洁[2] 黄 梅[2]

[1]哈佛大学东亚语言及文化系 马萨诸塞 波士顿
[2]北京语言大学 北京 100083

提要 文章以汉语书面语体理论"韵律语法"为基础,首先介绍了汉语书面语体的庄雅特征,如嵌偶词、合偶词、书面语句型等;其次给出了庄雅特征的量化方法及庄雅度自动测量的方法;再次通过 HSK(高等)考试作文的语料验证了庄雅度计算结果的可靠性;最后讨论了庄雅度自动测量技术的应用领域并指出其现实及潜在意义之所在。

关键词 书面正式语体 韵律语法 庄雅特征 庄雅度测量
中图分类号 H087 **文献标识码** A **文章编号** 1671 - 9484(2008)02 - 0113 - 14

1 引言

长期以来,"什么是汉语的书面语"、"书面语与口语有何不同"等问题,不仅困扰着汉语的语言学,同时也是文学、语文学及其他相关领域的大问题。口语和书面语的关系搞不清,书面语的性质就无从谈起。反之亦然:书面语的性质不定,它与口语的关系就不明。现代书面语的本质属性就在于它要区别于口语,或者说,它是近百年来使用者有意无意地使之与口语拉开距离的产物。(参冯胜利 2003)"拉开与口语的距离"是书面语的基本属性。显然,这与通常"书面语 = 白话文,白话文 = 口语"的看法和提法大相径庭。我们不反对书面语来源于口语的说法,但笃信书面语必须独立于口语而后可生。原因很简单,不如此则不能与口语拉开距离,而所谓"书面语",在我们的定义下必须严格地理解为"现代汉语书面正式语体"。(冯胜利 2006)根据这个定义,就不难理解为什么现代汉语书面语体需要而且必须"脱离"口语而独立的原因所在。不脱离口语、不与口语拉开距离,就显得不正式——语言的这种"语体社会功能"迫使"我手写我口"的白话文,中途分轨,生出一支必须"手不写口"的正式庄雅语体。这是汉语书面语研究上的一个重要突破口,因为不管语言学、文学,还是语文学,一直都把书面语和口语混而为一,[1] 而在五四运动"我手写我口"历史潮流中,至今未有提出书面语应当而且必须独立于口语的主张(恐有复辟之嫌也是原因之一)。[2] 因此,在"书口同一"的大潮下,似乎无论接见外宾还是跟家人吃饭,一套"对襟挽裆裤"便可敷用一切场合。事实当然不是这样,因此,文言文被打倒后,必然要生长出一种别于口语的正式语体(以前的正式语体是文言文)。为什么?"方圆异德",故"雅俗殊形矣"! 我们认为,现代书面

[收稿日期] 2007 年 11 月 27 日 [定稿日期] 2008 年 2 月 1 日
* 本文得到了《语言科学》匿名审稿专家提出的宝贵意见,在此表示感谢!

〔1〕 当然也有提出彼此不同者,如朱德熙(1987),胡明扬(1993)等。

〔2〕 唯季刚(黄侃)先生有说,谓:"常语趋新,文章循旧,方圆异德,故雅俗殊形矣"。(见《黄侃日记》199 页)

的正式语体既是源于口语的产物,又是脱离口语的结果。正因如此,无论语言学、文学,抑或语文学,均不能对写下来的文字"一视同仁",因为其中有口语语体,也有正式语体,而二者不仅功能不同(语用功能和社会场合),而且系统也不一样(词汇系统和语法系统)。

从"书面语"到"书面正式语体"的提法不仅是术语的不同、概念的变化,重要的在于对对象本质的揭示。如果说书面正式语体的本质在于"用词汇和语法的手段与口语系统拉开距离",那么研究和挖掘书面正式语体中的词汇和语法,则是书面语研究的又一大课题。汉语的书面语由哪些成分组成?什么样的语言特征使书面和口语呈现出语体的不同和差异?以往人们大多关注书面语与口语在词汇选择、句子长短方面的差异。这无疑是书面语与口语的不同特征,然而,仅仅如此仍不免隔靴搔痒,因为它没有触及到汉语书面正式语法的本质属性。

什么是汉语书面正式语法的本质属性呢?首先,如果从语法体系的角度来看书面语与口语的差异,那么现代汉语书面正式语体的语法体系就集中表现在它韵律语法的特点上,亦即韵律制约下的构词造句法。在近几年的工作中,我们研究了汉语书面语体的历史成因及运作模式;(冯胜利 2003,2005,2006)收集近乎全部的正式语体的表达形式,并编成《汉语书面用语初编》;(冯胜利 2006)与此同时,我们还进一步提出测量现代汉语文章"庄雅度"的可能和设想。这无疑开始改变原有对书面语只"抽象论说"而不能"据理操作"的研究状况。

然而,汉语文章的庄雅度是否真的可"测"?当然,依靠专家似乎可以,虽然其中不免带有一些主观的成分。那么能否用机器自动测量呢?这种想法在以前,若非异想天开,也属痴人说梦。然而,如下文所示,机器自动测量不再是一种理论的构想和实践奢望,因为它已成为一种可以具体实施的测量手段和技术。就是说,汉语书面语体庄雅度的自动测量可以说基本上从理论变成了现实。(参访风雅网站,即可窥见一斑:http://poem.guoxue.com:8080/)本文即以韵律语法的理论为框架,分别介绍如何实现现代汉语书面语庄雅度自动测量的构式和技术。第 2 节简单介绍韵律语法及现代汉语的书面语特征,第 3 节介绍庄雅度的计算方法,第 4 节讨论庄雅度的检验和应用,第 5 节指出本项研究的意义、目前的问题及进一步的工作。

2 韵律语法及现代汉语书面语特征

2.1 韵律语法体

人们说话时有的地方轻,有的地方重,有的音长,有的音短,这就是韵律。在汉语韵律的体系中,两个音节组成一个音步,一个音步实现一个韵律词。语言都有韵律,然而,如果该语言的韵律不但控制词汇,而且制约句法的运作,那么这个语言的语法就是韵律语法。其定义可表述如下:(参冯胜利 2006)

韵律语法

如果该语言的计算系统(Computational System)必须在韵律规定的条件下才能合法运作的话,那么这种语言的语法就是韵律语法,亦即韵律制约下的构词造句法。

i) $^{*}[\sigma] \rightarrow [\sigma\sigma]_{\text{韵律词}}$

单音节不足构成一个音步因此不成韵律词,故古语必双而后独立(如"果知");

ii) $[\sigma\sigma]_{\text{韵律词}} \rightarrow [\sigma\sigma]_{\text{韵律词}} + [\sigma\sigma]_{\text{韵律词}}$

韵律词必选韵律词与之搭配,故书面语"双必合双"而后上口(如"进行");

iii) 文章的内容越庄雅,韵律词的要求就越严格。

韵律语法是韵律制约下的构词造句法,主要体现在词法和句法两个方面。在构词句法模式中,"单+单"这种运作模式使得大量"耳听可懂"的文言单音词沿用至今,活跃在现代汉语书面语中,充当书面语体里面的庄雅功能,这是现代汉语书面语的特征之一(亦即"嵌偶词",见下文)。在句法运作模式为"双+双"的环境里,现代书面正式语体从自身的系统中发展出来一批双音词,它们的形态句法要求与它们的搭配成份也必须成双,造成"双音词+双音节"的搭配规则,以此呈现正式语体的庄重色彩,这是现代汉语书面语的特征之二。不难看出,"古语必双而后独立"的规则和"今语双音必接双音"的要求,本身都是新生的韵律语法,古代是绝对没有的。

2.2　嵌偶词

如上所述,现代汉语书面语中活跃着一批必须组单成双之后才能独立运用的单音节文言词,我们称之为"嵌偶单音词"(下文简称为"嵌偶词")。嵌偶词严格地遵循着"单+单"的组构方式,如果它不和另一个单音词(或者也是嵌偶词,或者不是)合成一个韵律词就不能合法出现。〔3〕根据我们的研究,目前已收集嵌偶词350个左右。表1给出了部分嵌偶词及其组双实例、误例。

嵌偶词	组双实例	误例
暗	暗查,暗送,暗想	暗检查,暗赠送,暗思考
备	备尝,备感,备受	备尝试,备感到,备经受
餐	订餐,送餐,三餐	预订餐,外送餐,三顿餐
错	错砍,错杀,错认	错砍伐,错杀死,错辨认
返	返京,返美,返校	返北京,返美国,返学校
广	广传,广交,广寻	广传播,广结交,广寻找
景	观景,选景,雪景	观赏景,选择景,白雪景
力	力保,力拒,力劝	力保证,力拒绝,力劝说
享	同享,独享,尽享	共同享,单独享,尽情享
资	闲资,添资,敛资	闲余资,添加资,收敛资

表1　嵌偶词举例

2.3　合偶词

上文说过,现代汉语书面语还自身发展出一批双音词,这类双音词文言中没有,口语中罕用,只在现代汉语书面语中出现,且必须和另一个双音词组成"双+双"的韵律模块后才能合法出现,这种双配双的双音词称之为"合偶双音词"(下文简称为"合偶词")。目前的研究已收集合偶词500个左右。下页表2给出了部分合偶词及其配双实例、误例。

2.4　书面语句型

除上述嵌偶词、合偶词之外,现代汉语书面语还使用大量的文言句型,这些句型在口语中不用,可称为"书面语句型"。这种句型在书面正式语体中同样扮演着正式、庄雅的角色。这是现代汉语书面语的特征之三。目前已收集书面语句型300个左右。下页表3给出了部分书面语句型及其文白对应实例。

〔3〕 关于嵌偶词的严格定义,参冯胜利(2006)序言。

合偶词	配双实例	误例
保卫	保卫人民,保卫祖国,保卫家乡	保卫人,保卫国,保卫家
从事	从事教学,从事写作,从事翻译	从事教,从事写,从事译
光临	光临寒舍,光临我校,光临大会	光临家,光临校,光临会
建筑	建筑桥梁,建筑公路,建筑房屋	建筑桥,建筑路,建筑房
合法	合法居住,合法买卖,合法分配	合法住,合法买,合法分
宏伟	宏伟气势,宏伟宫殿,宏伟蓝图	宏伟气,宏伟宫,宏伟图
极为	极为严格,极为深刻,极为灵验	极为严,极为深,极为灵
加以	加以评论,加以改正,加以补充	加以评,加以改,加以补
进行	进行调查,进行讨论,进行批判	进行查,进行论,进行批
消耗	消耗能源,消耗体力,消耗时间	消耗能,消耗力,消耗时

表 2　合偶词举例

书面语句型	例句(文)	例句(白)
中不乏	学生中不乏知识渊博的人。	学生中也有不少知识渊博的人。
称之为	《白马论》不可简单称之为诡辩。	不能简单地把《白马论》说成是诡辩。
所以然	哈佛有名,其所以然者在于大师云集。	哈佛大学很有名,其中的原因就在于学校里的大师特别多。
盖……之故	此盖款不敷出之故。	这大概是因为钱不够花的原因。
且……,何况……	他白话文且不能读,何况文言文呢?	他连白话文都不能读,还用说文言文吗?
如……,则……	如要发展,则必须改革。	要是想发展,就必须改革。
就……而言	这类文字,就诗学价值而言更显珍贵。	这类文字,从诗学价值上看,就更表现出它的珍贵。
将 v 矣	民营企业用人如果任人唯亲,企业将亡矣。	民营企业用人如果任人唯亲,企业就要灭亡了。
较……为 a	他晚年画作远较其早期作品为亲切。	他晚年的画作比他早期的作品更亲切。
借 v| n 以 v	文人常借喝酒以创造灵感。	文人常常用喝酒的方式来创造灵感。

表 3　书面语句型举例("v"代表动词,"a"代表形容词,"n"代表名词,"|"表示逻辑"或")

3　庄雅度的测量

我们将嵌偶词、合偶词及书面语句型作为体现现代汉语书面正式语体庄雅特征的主要语言成分,这些不同的庄雅特征不仅有理论的根据,同时也有数量极限(亦即有可穷尽性)。此外,还将收集到的现代正式语体中使用的典雅功能词(以文言为主)和 HSK(汉语水平考试)词汇大纲中的丁级词,作为庄雅成分的补充特征。当然,还有哪些形式可以作为鉴定庄雅特征的补充成份,仍在进一步的研究之中。仅就目前的成果而言,我们所选取的上述庄雅成分(或庄雅特征)不仅可以表现文章的庄雅度,而且比较准确地反映出作文水平的高低等差。最有意义的是,有了可识别的特征(庄雅成分),便可以设计一套计算方法,从而对现代汉语书面语的庄雅度进行自动测量。〔4〕

〔4〕需要特别指出的是,本文的庄雅度自动测量,和目前国外有些"书面语和口语特征的统计对比"很不一样:前者具有写作水平测试的功能(见下文),后者(至少现在)还没有可能制成软件自动测量文章等级的结果。

下面先看嵌偶词。如何计算嵌偶词在一篇具体文章中所占的比例呢？我们以嵌偶词的数量为分子，分为次数(token)和种数(type)两种情况(其他特征也都分为这两种情况)，以文章的字数为分母。见公式1、公式2：

公式1　嵌偶词比例(token) = 嵌偶词的出现次数(token) / 字数

公式2　嵌偶词比例(type) = 嵌偶词的出现种数(type) / 字数

在计算合偶词在一篇具体文章中所占的比例时，我们以合偶词的数量为分子，以文章的词数为分母。见公式3、公式4：

公式3　合偶词比例(token) = 合偶词的出现次数(token) / 词数

公式4　合偶词比例(type) = 合偶词的出现种数(type) / 词数

在计算书面语句型在一篇具体文章中所占的比例时，我们以书面语句型的数量为分子，以文章的句数为分母。见公式5、公式6：

公式5　书面语句型比例(token) = 书面语句型的出现次数(token) / 句数

公式6　书面语句型比例(type) = 书面语句型的出现种数(type) / 句数

在计算HSK丁级词在一篇具体文章中所占的比例时，我们以丁级词的数量为分子，以文章的词数为分母。见公式7、公式8：

公式7　HSK丁级词比例(token) = HSK丁级词的出现次数(token) / 词数

公式8　HSK丁级词比例(type) = HSK丁级词的出现种数(type) / 词数

在计算典雅功能词在一篇具体文章中所占的比例时，我们以典雅功能词的数量为分子，以文章的词数为分母。见公式9、公式10：

公式9　古汉语功能词比例(token) = 古汉语功能词的出现次数(token) / 词数

公式10　古汉语功能词比例(type) = 古汉语功能词的出现种数(type) / 词数

由于在计算上述五种特征比例时所采用的分母并不一致，庄雅度的计算不宜将各特征比例直接累加，因此我们采用对各特征比例进行加权平均的办法来计算庄雅度；又由于各特征在体现庄雅度时究竟孰轻孰重这一问题还有待研究，(这一项目正在进行)因此我们暂时决定对各特征比例的权重取20%，即采取算数平均法来计算庄雅度，〔5〕见公式11：(也分token和type两种情况，在公式表述上不再区分)

公式11　庄雅度 = (嵌偶词比例 + 合偶词比例 + 书面语句型比例
+ HSK丁级词比例 + 古汉语功能词比例) *20%

4　庄雅度的检验与应用

由上可见，一旦实现了庄雅成分特征的量化，庄雅度的自动测量就可以把理论转化成可以操作的算式。从理论上说，单纯实现书面语特征的检索及庄雅度的自动测量并不是最终的目的。我们面临的最终问题是：计算出来的庄雅度究竟有多大程度的可靠性；它的可靠性又能带来它的多大程度的实用性。因此，当完成了“书面语的特性是什么”、“书面语的正式特征有哪些”、“书面语正式特征的量化方式”，以及“庄雅度的自动测量的软件编程”的任务以后，我们面临的挑战就不再是“能不能测量的问题”，而是测

〔5〕毫无疑问，上述五个特征的权重比例虽需理论上进一步论证，但如下文所示，仅从它们各占20%的比例上，就可以看出不同特征所反映的不同典雅度。同时，这种典雅度的不同结果，对文章的测试具有重要的意义。

量出来的结果(或数字)到底可靠与否、到底有用与否的问题。显然,我们现在面对的是一个严峻的庄雅度自动测量的检验问题。

怎么解决这个问题呢?当然,我们可以随便拿一篇文章来测试,然后看其结果如何。然而,测出的结果用什么标准来鉴定呢?我们知道,诗无达诂,自古而然,更何况一篇文章的庄雅度(或好坏)本来就见仁见智,公婆不一。单篇检验不是不行,只是可靠度不够。那么如何测定我们"庄雅度自动测量"的技术呢?经过多番实验和研究,我们最终选用了 HSK(高等)考试作文的语料进行实验。我们以 2001 - 2005 年 HSK 作文 3949 篇为实验样本(经过 HSK 专家评定的等级)进行测试实验,结果表明:我们设计的庄雅度与专家评定的作文等级,彼此对应,基本准确。这就意味着庄雅度确与文章的优劣直接相关。下面就详细介绍实验方式和结果。

首先,我们把近四千篇 2001 - 2005 年 HSK 作文作为实验样本。其中作文分数采用了两种等级:分数等级 1 为 A、B、C、D 四等,依次由高到低,C 为及格;分数等级 2 为十二等,其中 65 分为及格。两种等级的对应见表 4:

分数等级 1	A	B	C	D
分数等级 2	95 、90 、85	80 、75	70 、65	60、55 、50 、45 、40

表 4　分数等级对应

各分数等级分别取庄雅度的平均值。试验的结果正如我们理论所预测的:分数由高到及格线时(A - B - C 或 95 - 65),无论是整体的庄雅度还是各分项(嵌偶、合偶、句型、功能词、丁级词)都与人工评分的分数的等级基本上成正比。

当然,分数在及格线以下时,庄雅度与分数之间就没有什么必然联系了,有时反而会偏高。其主要原因是 HSK 作文对字数有要求(400 字),当篇幅很短时就不能及格了。而篇幅短的话,无论作文实际的庄雅度是高是低,由于计算庄雅度时分母相对会小,结果都会偏高。分数与平均字数的对应关系见表 5、表 6。分数与庄雅度及各分项的对应关系见表 7 —表 18 以及相应的柱形图。在我们的实验中,合偶词部分不太成比例,原因待考察。

	A	B	C	D
平均字数	455	431	390	315
作文数量	470	868	1687	924

表 5　字数-分数等级 1

	A	B	C	D
token(%)	1.95	1.61	1.43	1.44
type(%)	1.60	1.32	1.17	1.19

表 7　庄雅度-分数等级 1

	95	90	85	80	75	70	65	60	55	50	45	40
平均字数	472	468	444	441	424	407	378	334	294	265	236	224
作文数量	54	162	254	345	523	756	931	617	186	80	30	11

表 6　字数-分数等级 2

因此,除去 D 级(平均字数 315)不计,前三等级的预测准确度,可从表 7 所示的庄雅度的分布等级清楚地看出来(下页图 1 为其柱形图)。换言之,人工评分(A、B、C 级)和庄雅度的对应关系为:

人工评分的结果　　机器测量的结果

A 级　=　1.95 %

B 级　=　1.61 %

C 级　=　1.43 %

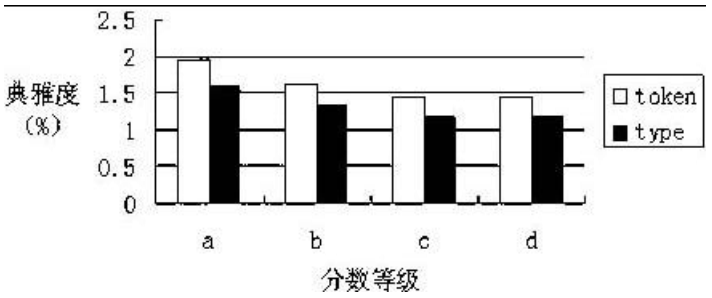

图 1　庄雅度-分数等级 1

A、B、C 三级的等差已经相当清楚,这是总体测量的结果。如果我们分项来看,那么其中等级差别的细节就更清楚了。先看嵌偶词与人工评分的对应性:

	A	B	C	D
Token(%)	0.43	0.30	0.25	0.24
type(%)	0.35	0.25	0.21	0.20

表 8　嵌偶-分数等级 1

表 8 数据说明,在 HSK 的人工评分结果里,获得 A 级(85 - 95 分)的作文,在自动测量结果里的庄雅度为 0.43 %,B 级(75 - 80 分)在自动测量结果里的庄雅度为 0.30 %,C 级(65 - 70 分)的在自动测量结果里的是 0.25 %。这种明显的等级差别,可以通过图 2 清楚地看出来:

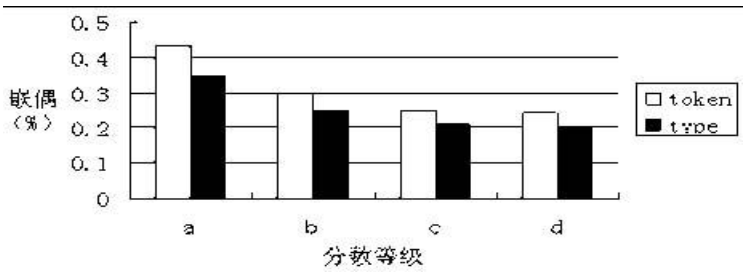

图 2　嵌偶-分数等级 1

毋庸讳言,嵌偶词使用量的多少不仅反映了评分者的对文章等级的判断,而且反映了作者写作水平的高低。试验的结果和我们的预测不谋而合。下面看表 9 的合偶词的分布情况:(实例见表 2)

	A	B	C	D
Token(%)	2.01	2.09	1.91	1.91
type(%)	1.68	1.67	1.52	1.57

表 9　合偶-分数等级 1

合偶词似不能反映 HSK 作文评分后的等级的不同,因为 B 级分数高于 A 级。如图 3 所示:

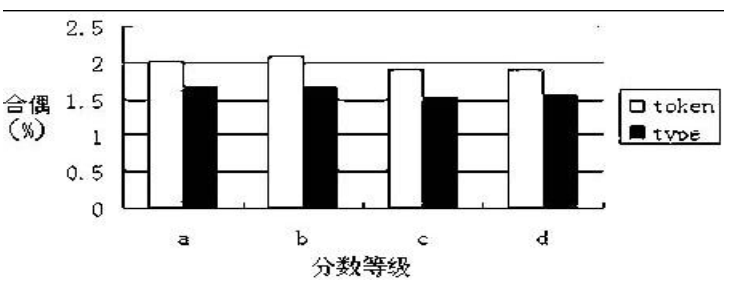

图 3　合偶-分数等级 1

不仅三个等级没有很大的区别,而且该低的不低、该高的不高。什么原因使得书面语自身系统中的"双+双"形式在作文水平上泯然无别呢?显然,这是一个有待深入研究的重要课题,其中的分析,还需另文专述。下面表 10 是书面句型分布的情况:

	A	B	C	D
token(%)	1.98	1.53	1.27	1.28
Type(%)	1.84	1.38	1.16	1.14

表 10　句型-分数等级 1

从上表可见,书面语句型的等级分布,非常理想:1.98 % > 1.53 % > 1.27 %。其中每级均呈递加增长趋势,B 级以 0.26 %的增值量高于 C 级,而 A 级则以 0.45 %的增值量高于 B 级。这种递加式增值量可从图 4 清楚地看出来:

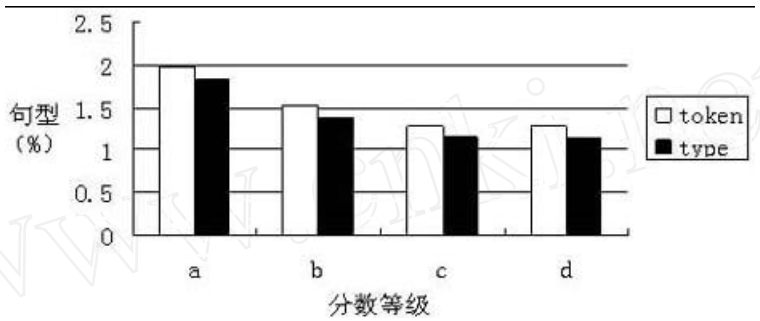

图 4　句型-分数等级 1

句型的分布无疑可以帮助我们测定庄雅度。那么功能词的结果如何呢?请看表 11:

	A	B	C	D
Token(%)	1.10	0.89	0.72	0.72
type(%)	0.83	0.67	0.55	0.57

表 11　功能词-分数等级 1

显然,功能词是以平均 0.19 %的比差率逐级增值,从而将 A、B、C 三个等级拉开距离。如图 5 所示:

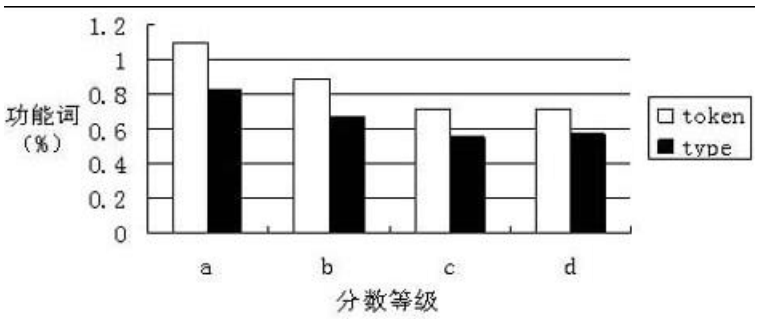

图 5　功能词-分数等级 1

上文说过,为了发掘汉语书面语体的正式特征,我们选用丁级词作为测量的成分。结果请看表 12:

	A	B	C	D
token(%)	4.21	3.25	3.03	3.09
type(%)	3.31	2.63	2.42	2.46

表 12　丁级词-分数等级 1

很明显,丁级词同样具有区别作文等级的功能:ABC 三个等级分别以 4.21 % > 3.25 % > 3.03 % 等

级差异彼此区别;其平均比差率为 0.59 %。图 6 可见其详:

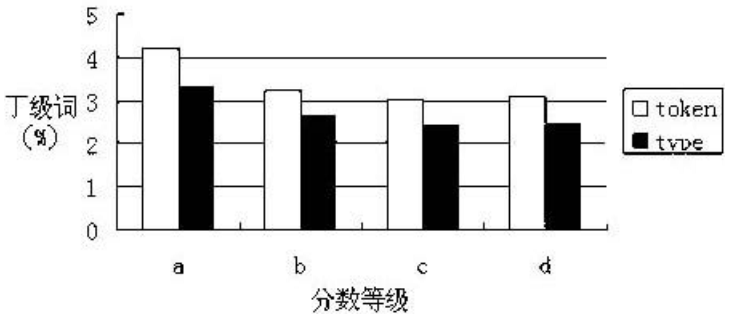

图 6　丁级词-分数等级 1

分析至此,除合偶词一类以外,其他四类正式语体的特征均能很好地反映 HSK 人工评分的等级。当然,上面看到的是我们将近 4000 篇作文分为 ABCD 四个等级后,检验它们和庄雅度测量结果的"对应性"。尽管就大类而言,人机测量的结果具有高度的对应性,我们仍然不知道在评分的细节等级上(如 95、96、97 ……),人机的对应性是否存在。毋庸讳言,在作文的评分等级上,即使是专家,即使再精审,也难免"大类有别,细目难分"。对作文成绩而言,虽然 80 分和 90 分或有天壤之别,90 分和 91 分就很难说谁好谁坏。因此,我们舍细目而验小类:以 5 分为界,来检验庄雅度自动测量软件和专家评分的对应性。先看表 13 的总体的情况:

	95	90	85	80	75	70	65	60	55	50	45	40
token(%)	2.57	1.91	1.84	1.70	1.55	1.41	1.46	1.50	1.31	1.38	1.43	1.74
Type(%)	2.13	1.58	1.50	1.36	1.29	1.15	1.18	1.21	1.07	1.20	1.28	1.42

表 13　庄雅度-分数等级 2

从 70 分以上,我们"庄雅度自动测量器"得出的结果,和人工评分的等级,基本上相互吻合。然而,在 60 和 70 之间,则情况复杂起来。请看图 7:

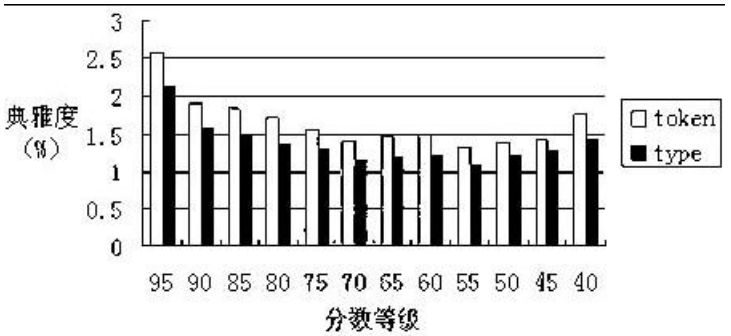

图 7　庄雅度-分数等级 2

问题出在哪里了呢?为什么 60 和 65 分两级的庄雅度居然高于 70 分呢?值得将来深入研究。我们考虑,这可能不仅是测量手段的问题,而且还牵涉到语言习得的阶段问题。注意,这 60 - 70 之间的分数,正好反映了中文习得过程中的"中级阶段"的中文水平。而上面 ABC 三级分数和庄雅度的对应性,所以整齐、严格的原因之一,就是我们把分数和庄雅度不对称的"65 - 70"作为一个级别(或整体)来处理的缘故。在八级的系统里(从 60,65,70 ……95 共八级),原来的一个等级,分成了三个等级,其中的问题也就显露出来。问题在于,是不是在所有的特征成分分布里,60 - 70 之间都具有同样的特殊性呢?下面我们分别检验。先看表 14 嵌偶词的情况:

	95	90	85	80	75	70	65	60	55	50	45	40
token(%)	0.48	0.46	0.40	0.30	0.30	0.25	0.26	0.25	0.24	0.16	0.22	0.22
type(%)	0.39	0.38	0.33	0.25	0.25	0.21	0.21	0.21	0.20	0.11	0.22	0.22

表 14　嵌偶-分数等级 2

由上可见,虽然 80 和 75 之间没有区别,但是 75 和 70 之间确有距离。有意思的是,60 - 70 之间没有等级之差。图 8 更形象地反映出这一点:

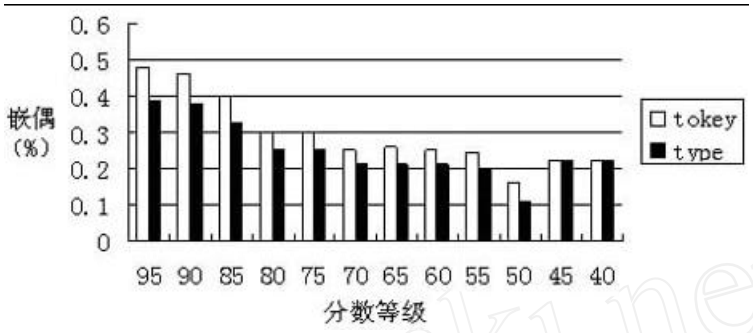

图 8　嵌偶-分数等级 2

嵌偶词是这样(我们称之为"60 - 70 分的庄雅度效应"或"中级效应"),那么合偶词是什么情况呢?请看表 15:

	95	90	85	80	75	70	65	60	55	50	45	40
token(%)	2.63	1.88	1.97	2.15	2.05	1.93	1.89	1.97	1.80	1.80	1.96	1.60
Type(%)	2.20	1.59	1.63	1.69	1.65	1.53	1.51	1.60	1.47	1.54	1.72	1.51

表 15　合偶-分数等级 2

和上文(表 9 及图 3)所反映的情况一样,合偶词需要进一步研究。因为它们的分布似乎无视作文的等级,没有什么规律可言。请看图 9:

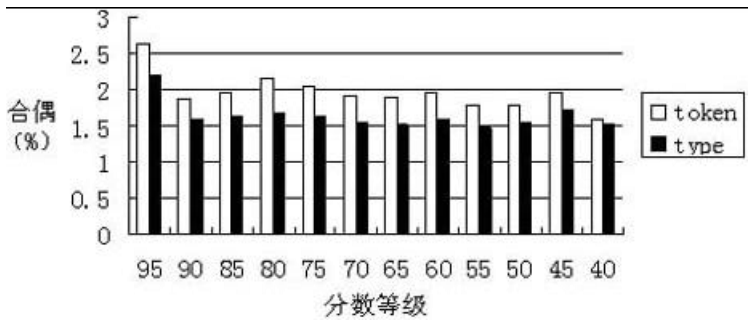

图 9　合偶-分数等级 2

"60 - 70 分的庄雅度效应"(或教学法上的"中级效应")在"书面句型"里的情况,比在嵌偶词里还突出。请看表 16:

	95	90	85	80	75	70	65	60	55	50	45	40
token(%)	3.26	1.75	1.86	1.72	1.40	1.17	1.34	1.42	0.80	1.29	1	2.76
type(%)	2.79	1.69	1.73	1.51	1.29	1.08	1.22	1.23	0.76	1.29	1	1.75

表 16　句型-分数等级 2

不难看出,从 70 - 95 之间,庄雅度和评分等级基本对应(基本没有越级的现象)。然而,在 60 - 70 之间,情况不同了,70 的庄雅度低于 65,65 低于 60,造成参差不齐的分布局面,如下页图 10 所示:

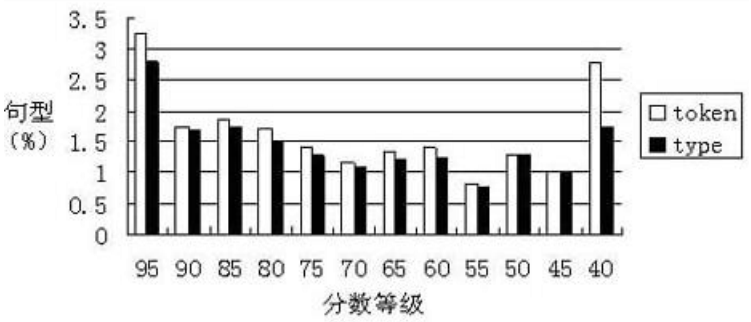

图 10　句型-分数等级 2

我们再来看表 17 的功能词。无独有偶,功能词也表现出一种明显的"60 - 70 分的庄雅度效应":

	95	90	85	80	75	70	65	60	55	50	45	40
token(%)	1.55	1.03	1.04	0.91	0.87	0.68	0.75	0.74	0.63	0.63	1.14	0.73
type(%)	1.17	0.79	0.79	0.67	0.66	0.53	0.56	0.57	0.51	0.56	0.91	0.73

表 17　功能词-分数等级 2

和书面句型差不多,70 分的庄雅度低于 65 和 60。所以也造成分布图的高低不平,如图 11 所示:

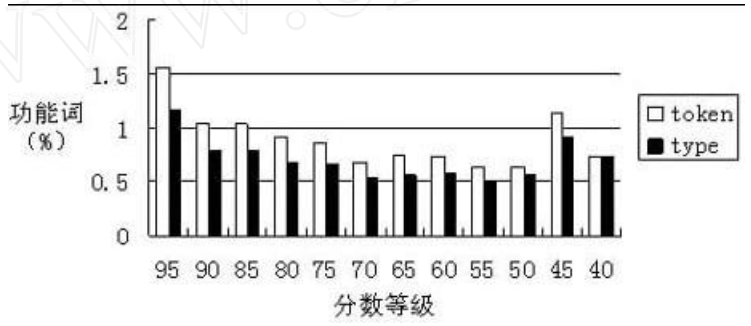

图 11　功能词-分数等级 2

最后,再看表 18 丁级词的情况:

	95	90	85	80	75	70	65	60	55	50	45	40
Token(%)	4.94	4.44	3.92	3.41	3.15	3.02	3.03	3.10	3.09	3.01	2.83	3.39
Type(%)	4.08	3.46	3.04	2.70	2.58	2.42	2.43	2.47	2.39	2.50	2.57	2.90

表 18　丁级词-分数等级 2

表 18 告诉我们,70 的庄雅度低于 65,65 的庄雅度低于 60,如图 12 所示:

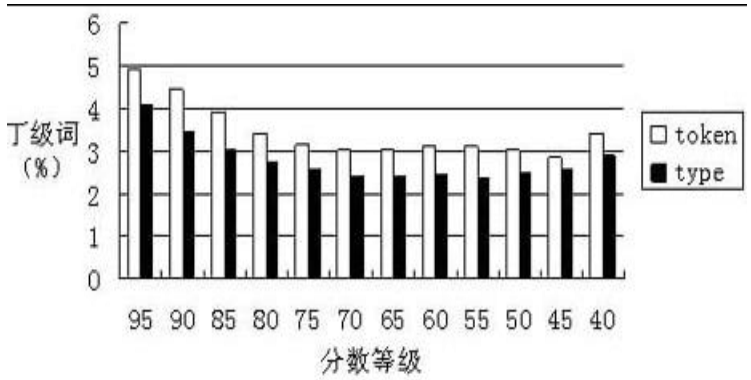

图 12　丁级词-分数等级 2

不难看出,“60 - 70 分的庄雅度效应”在丁级词里也表现得非常突出。是中级程度无庄雅呢?还是庄雅度测量手段不全面呢?无论是什么原因,本文的结果都给我们将来的研究提出了一个重要的问题。无论我们如何看待“60 - 70 分的庄雅度效应”,这种庄雅效应本身就暗示着本项研究的重要意义,如果除去“60 - 70 分的庄雅度效应”而不论(或把它们作为一类)的话,那么我们研制出的“庄雅度自动测量”技术,基本可以反映出作文的等级。反过来说,如果庄雅度测量在 70 - 90 的分域里基本精确无误的话,那么 60 - 70 分域里的问题则另有原因。其中的奥秘,值得进一步发掘和研究。

以上我们验证了庄雅度与 HSK 作文等级的相关性,同时我们也认识到庄雅度并不能作为测量作文等级的单一指标。

如果各类作文的样本分布都集中在接近各自庄雅度平均值的区间内,各类之间相对独立(即没有交叉或少有交叉)的话,那么庄雅度就可以作为理想的单一指标测出作文的不同等级。理想的分布状态如图 13 所示,横轴表示庄雅度,纵轴表示作文数量。在理想分布的情况下,就可以设定各分数等级对应的庄雅度的阈值,即 a ≥x3,x3 > b ≥x2,x2 > c ≥x1,x1 > d,从而单纯根据庄雅度来进行自动评分。

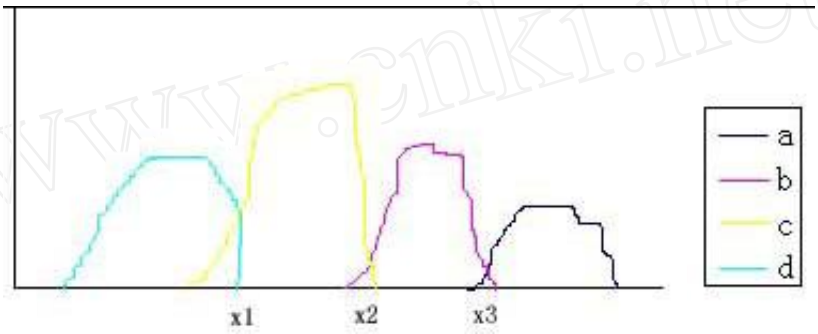

图 13　丁级词-分数等级 2

然而,通过对数据的考察,我们发现各类作文的真实分布呈交叉重叠状态,如图 14 所示:

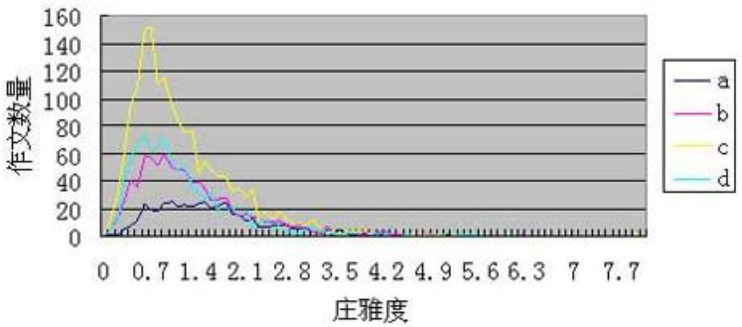

图 14　abcd 四类作文分布

图 14 以庄雅度(type)为横轴,以作文数量为纵轴。由此可知,庄雅度虽然是测量作文的参数指标,但是它绝不能作为测量作文等级的单一指标。〔6〕

5　余论

本文以近年发展的汉语书面语体理论为基础,第一次初步实现了庄雅度的自动测量。不仅如此,还

〔6〕 还有哪些指标与写作水平相关,以及使用何种方法将各类指标综合利用可以让计算机给出一个合理的分数是作文自动评分领域需研究的内容,本文暂不讨论。

设想和实验了其应用的领域。当然,目前的工作还不够细致,无论从理论研究还是从工程实现的角度,都有待进一步的深入。譬如:

1)特征的进一步挖掘。除本文提到的五类特征外,还有哪些语言特征能够反映书面语的庄雅度是有待研究的问题。

2)现有特征的完善。其五类特征的实例数量都是相对有限的,还需进一步穷尽式收集实例以求完善。此外,各特征的界限存在交叉,如"逃避"既是合偶词,又是 HSK 丁级词,那么遇到这种情况时是允许身兼多职(目前的做法)还是将之归为一类也是仍需研究的问题。

然而,无论存留问题怎样,这里的实验效果,已经证明我们有关书面语体理论的正确性。理论的价值更在于实用。实践证明,我们的检索方案和定量分析,具有广泛的实用价值。庄雅度的自动检测程序,不仅可以用来教,而且可以帮助学。〔7〕更重要的是,正如冯胜利(2005)所预测的那样,该测量系统还可以给"汉语的文白规范、文章难度的等级测定(readability)、汉语写作等级的鉴定,高级汉语的教学内容、高级汉语教材的编写以及高级汉语的标准"等紧迫问题的解决,提供一个客观的标准和突破的门径。

毋庸讳言,这里开发的自动检索技术与定量分析方法,在作文测试的研究领域,还是第一次。其技术成果可进而用于区分不同文章的文体,因为文体的不同和文章庄雅的等级息息相关。我们认为:庄雅度、难易度、优劣度等方面的等级测定,都是"牵发动身"、"此起彼伏"的对应关系。庄雅度与文章语体的相关性不言而喻,而与难易度的相关性主要是语言的表达而不涉内容的深浅,这也无需多言。庄雅度与文章优劣的相关性,则是通过作者语言水平的高低来表现、来测量的。

毫无疑问,区分文章语体的庄雅度、难易度、优劣度,可直接用于汉语教材的编写、汉语考试试题的编写,以及写作的评分等诸多领域。区分文章的语体主要指区分书面语体的文章和口语体的文章,这对编写不同课型的对外汉语教材来说,有着极其重要的指导意义——口语教材选入的文章不宜书面语色彩太重。区分文章的难易度,对教材的编写和考试的试题,也有直接的帮助,因为不同年级的教材其所入选的课文的难易度应该有所差别,不同等级水平的汉语考试,其听力、阅读等题型的相关文章,也应有难易度上的等级差别。区分文章的优劣则对评价学习者的汉语写作水平意义重大。

总而言之,这里的科研成果,首次为"作文机器评分"打开了一个窗口。我们希望在不断开发和完善的道路上,使这项技术不仅为汉语的研究与教学,而且为其他相关领域的需求,发挥其更大的作用。

参考文献

冯胜利　2003　书面语语法与教学的相对独立性,《语言教学与研究》第 2 期,53 - 63 页。

冯胜利　2005　《汉语韵律语法研究》,北京:北京大学出版社。

冯胜利　2006　《汉语书面用语初编》,北京:北京语言大学出版社。

胡明扬　1993　语体与语法,《汉语学习》第 2 期,1 - 3 页。

黄　侃　2001　《黄侃日记》,南京:江苏教育出版社。

张正生　2005　书面语定义及教学问题初探,载冯胜利、胡文泽编,《对外汉语书面语教学与研究的最新发展》,332 - 338 页,北京:北京语言大学出版社。

朱德熙　1987　现代汉语语法研究的对象是什么,《中国语文》第 5 期,321 - 329 页。

〔7〕 这里指对外汉语的教与学,因为本文试验的最佳效果,只限于把汉语当作外语的作文测试。

作者简介

冯胜利,笔名冯利,男,1955 年 5 月 15 日生于北京。1977 年考入北京师范大学历史系,1979 年考取北京师范大学中文系古汉语专业研究生,从陆宗达先生治《说文》,1982 年毕业留校。1986 年赴美国宾西法尼亚大学语言学系读书,1995 年获博士学位。1994 - 2003 年在堪萨斯大学东亚系执教,任副教授。2003 年至今在哈佛大学东亚系执教,任教授及中文部主任,同时兼任北京语言大学长江学者、讲座教授。研究兴趣在韵律构词学及韵律句法学的建立以及历史句法学和对外汉语教学。出版专著有《汉语韵律句法学》等,并在《中国社会科学》、《中国语文》和 *Linguistics*、*East Asian Linguistics* 等杂志发表论文数十篇。

王洁,女,1980 年 2 月生,山东青岛人。北京语言大学语言信息处理研究所博士研究生,研究方向为自然语言处理。

黄梅,女,1979 年 11 月生,河北唐山人。北京语言大学博士研究生,研究方向为韵律句法学。

An Automatic Feature Checking Algorithm for Degree of Formalities in Written Chinese

Feng Shengli[1,2] Wang Jie[2] Huang Mei[2]

[1] *Department of EALC, Harvard University, Boston Massachusetts USA*
[2] *Beijing Language and Culture University, Beijing* 100083

Abstract Based on Prosodic Grammar, this paper introduces the formal features of written Chinese, including 1) monosyllabic words used in disyllabic templates, 2) disyllabic words used in disyllabic copulates, and 3) formal patterns in written Chinese. Secondly, an automatic feature checking algorithm is proposed for a quantitative analysis of the formalities in Chinese formal styles. Thirdly, the algorithm proposed in section 2 is verified by using nearly 4000 compositions from HSK (Hanyu Shuiping Kaoshi, or Chinese Proficiency Test), resulting in a precise matching between the degree of formalities calculated by the algorithm and the levels of HSK exam. Finally, it is argued for the first time that the automatic feature checking technology, can be used in a wide range of related fields, such as formality measuring, composition testing, readability scaling, style gradating, textbook compiling, L2 learning, literacy acquiring, and so on.

Keywords formal written Language prosodic grammar formal features formality measuring