《信息处理系统语言文字评测规范（草案）》三个规范研制报告

冯志伟 执笔

一 研制背景

（一）必要性、目的、意义和作用

信息处理系统是基于计算机技术、网络互联技术、现代通讯技术和各种软件技术，提供信息服务的人机系统，是由人和计算机等共同组成的能进行信息的收集、传输、分析、加工、处理、存储、更新和维护的系统。语言文字是信息的主要载体，语言文字信息处理系统的评测对于信息处理系统的研制和开发，具有重要的意义和作用。通过语言文字信息处理系统的评测有助于推进信息处理系统的进一步发展。

（二）社会和经济效益

随着信息技术的高速发展，信息处理系统已经成为社会语言文字传播和应用的主要工具之一，如何评价常见、典型的信息处理系统的语言文字处理水平和规范化情况是计算机专家和语言学家共同的关注点，也是国家语言文字工作主管部门所关心的问题和系统研制和开发部门特别关注的问题，对于我国的现代化和信息化，具有重要的社会和经济效益。

（三）立项情况

《信息处理系统语言文字评测规范》（草案）序列软规范是国家语委语言文字十五科研重点项目，由教育部语言文字应用研究所承担研制任务。研制目的是给出一个自然语言信息处理系统（例如，语音合成和文语转换系统、语音识别系统、机器翻译系统、语料库系统等）的语言文字评测规范，可供自然语言信息处理系统的语言文字评测以及有关的管理工作参考使用。由于《信息处理系统语言文字评测规范》包含的内容非常广泛，经过征求专家意见。目前我们先制订三个规范（草案）：《文语转换和语音识别系统语言文字评测规范（草案）》《机器翻译系统语言文字评测规范（草案）》《语料库系统语言文字评测规范（草案）》，作为《信息处理系统语言文字评测规范（草案）》序列软规范的一个部分。

二　研制过程

规范的研制过程主要包括对当前主要汉语信息处理系统的调研和对已有信息处理系统评测方法的研究，在调查和研究的基础上，课题解决了本规范研制遇到的诸多问题，初步形成了《信息处理系统语言文字评测规范》（征求意见稿），在征求意见稿的基础上，根据专家意见，我们又把这个征求意见稿分解成三个软规范：《文语转换和语音识别系统语言文字评测规范》（草案）《机器翻译系统语言文字评测规范》（草案）《语料库系统语言文字评测规范》（草案）。2006 年 8 月 20 日，教育部语言文字信息管理司组织专家对于课题组提交的三个规范草案进行鉴定，并且通过了鉴定。课题组根据鉴定会上专家提的意见又进行了进一步的修改，并在更加广泛的范围内征求业内专家的意见，历时两年多，最后形成了现在的三个规范文本。

（一）研制

1．信息处理系统及其相关评测的调研

2003 年项目立项后，课题组对国内外的语音合成和文语转换、语音识别、机器翻译、语料库等自然语言信息处理系统及其开展评测的情况进行了深入的调查研究，在调查研究中，我们特别认真地考察了如下的研究成果：

1）国家 863 智能接口评测的原理、方法、技术和内容的研究成果；

2）我国在语音信号处理系统评测中使用的语音测听方法和语音清晰度诊断性压韵测试法（Diagnostic Rhyme Test，DRT）的研究成果；

3）我国在军用通讯系统评测中使用的音质平均评价分（Mean Opinion Score，MOS）测试法的研究成果；

4）我国在广播节目声音质量评测中使用的主观评价方法和技术指标要求的研究成果；

5）北京大学计算语言学研究所俞士汶教授关于机器翻译评测的研究成果；

6）山西大学计算机系刘开瑛教授和东北大学计算机系姚天顺教授关于汉语文本切分和标注评测的研究成果；

7）华中师范大学计算机系何婷婷博士关于语料库元数据标准的研究成果；

8）国外在机器翻译评测中提出的 BLEU 和 NIST 等自动评测方法以及国外在自然语言信息处理评测研究方面的有关外文文献（参看参考文献 34-34）。

上述研究成果，特别是在国家 863 智能接口评测多年来积累的研究成果为我们制定本规范起到了重要作用。

2．七个调研报告的形成

根据上述的调查研究工作，我们形成了如下七个调研报告：

1）中国语料库研究调查报告；

2）关于机器翻译系统消歧功能测试的报告；

3）关于机器翻译现状和问题的报告；

4）关于机器翻译评测的调研报告；

5）关于文语转换和语音合成评测的调研报告；

6）关于语音识别评测的调研报告；

7）关于语料库评测的调研报告。

3．在软规范研制中需要思考和解决的几个典型问题

1）黑箱评测与白箱评测

a. 黑箱评测（black box assessment）：评测时不关心自然语言信息处理系统的内部机制和组成结构，主要根据系统的输入数据与输出结果进行判断。黑箱评测有助于了解信息处理系统外在的总体性能，又叫做“外在评测”（extrinsic assessment）。

b. 白箱评测（glass box assessment）；评测时需要对于自然语言信息处理系统的内部机制分别进行分析，逐一评测系统的各个组成部分的性能。白箱评测可以针对信息处理系统的各个组成部分分别进行，对于不同的部分准备不同的测试数据，从而判断所出现的错误是在那一个部分造成的，这样就可以为规则的调整和算法的改进提供可靠的数据。白箱评测有助于了解信息处理系统内部组成部分的性能，又叫做“内在评测”（intrinsic assessment）。

考虑到自然语言信息处理系统的语言文字评测基本上只涉及系统的外在的总体性能，尤其是系统输入和输出的语言文字水平及其规范化情况，因此主要采用黑箱评测方式。

2）系统评测与语言文字评测

系统评测主要是指对信息处理系统的性能进行评价。语言文字评测主要是依据语言文字的技术指标体系和有关规范，采用一定的方法和程序，对于自然语言信息处理系统及其组成要素的功能、特性和运行效果进行评价和检测。

因为在语音合成与识别、机器翻译等信息处理系统中，系统的性能很大程度上反映在语言文字的处理水平中，因此语言文字评测系统与系统评测是相互交叉在一起的，语言文字评测必须要包含系统性能评测的一些内容。863 智能接口评测已经开展多年，本规范涉及的系统性能评测主要参考了 863 智能接口评测大纲中的研究成果。

语言文字评测以国家已经发布的语言文字规范标准为基础，检查信息处理系统中与语言文字有关的用户界面（user interface, UI）、输入输出（Input/Output, I/O）等以及信息处理系统中以语言文字（如，文本的或语音的）作为载体的内容是否存在与语言文字

规范标准不一致或有冲突的内容，并对系统的语言文字处理能力、水平、规范化情况做出评价。

3）机器自动评测和人工评测

信息处理系统语言文字评测采用机器自动评测还是专家人工评测的方式是课题组需要思考的一个重要问题。课题组对已经发布的语言文字规范标准做了具体的量化分析，根据分析结果决定采取机器自动评测和人工评测相结合的方法。

已经发布的语言文字规范标准中，大部分是主观性较强的软性规范，机器自动评测缺乏一个客观规范。经过分析，也有部分规范标准可以全部或部分进行机器自动评测，如

序号	规范或标准
1	《第一批异体字整理表》
2	《部分计量单位名称统一用字表》
3	《简化字总表》
4	《GB/T 15834－1995 标点符号用法》
5	《GB/T 15835－1995 出版物上数字用法的规定》
6	《汉语拼音方案》
7	《中国人名汉语拼音字母拼写法》
8	《普通话异读词审音表》
9	《中文书刊名称汉语拼音拼写法》
10	《汉语拼音正词法基本规则》

课题组对每个语言文字规范都做了用于规范性检查和评测的具体分析，如：

《普通话异读词审音表》：该标准面向文教、出版、广播等部门及行业，用以规范普通话中有异读的词和有异读的作为“语素”的字的读音。《普通话异读词审音表》按字音的开首字母顺序排列，先为字形，然后是读音，之后附加说明，包括统读、文白读、词例、义项、应用范围等。

·检查目标：《普通话异读词审音表》是针对（字）音的语言文字规范标准，在本评测系统中，用以对语音识别及合成等信息处理系统中普通话异读词的读音、标音情况进行评测，检查其与国家规范标准相符合的程度。

·检查步骤：目前能够实施的有效检查是《普通话异读词审音表》中明确列举的词例以及注明“统读”的字的读音，对于表中那些被注明文白读或义项的字的读音只能进行有限制的检查，而对于表中没有明确说明的字的读音由于缺乏依据暂时无法进行检查。评测的具体规则如下：

①凡表中注明“统读”的字音，该字无论出现在任何词语环境中都应采用该读音，如出现其他读音为不规范。

②凡读音后注明词例的，该字在所出现的词语中应采用规定读音，如采用其他读音，为不规范。

③对于读音后注明文读的（表中用“（文）”表示）情况，有词例的，按词例进行检查；没有词例的，酌情进行检查，即如果受检内容明确为书面语言或文言成语，应采用该读音；如果不能明确受检内容的性质，则暂不予以检查。

④对于读音后注明白读的（表中用“（语）”表示）情况，有词例的，按词例进行检查；没有词例的，酌情进行检查，即如果受检内容明确为口语，应采用该读音；如果不能明确受检内容的性质，则暂不予以检查。

⑤对于读音后注明义项的，如果有词例，按词例进行检查；如果没有词例，酌情进行检查，即如果受检内容有义项标注，则按义项施行检查；如果受检内容没有义项标注，则暂不予以检查。

⑥若读音后注明“除××（较少的词）念乙音外，其他都念甲音”，则列举的词例应采用乙音，未列举的所有词语应采用甲音，如有违反，为不规范。

⑦有特殊说明或用于特殊研究目的的字音，不在上述检查的范围内。

⑧表中未作明确规定的轻声词的读音不受上述限制。

⑨除上述情况外，对于表中未作明确规定的字音，暂不予以检查。

除了软性规范外，一些语言文字水平如得体性、表达水平等也很难通过客观规范得到评测。

因此，信息处理系统语言文字评测采用机器自动评测和人工评测相结合的方法，规范中既包含机器自动评测的内容也包含专家人工评测的内容，具体实施评测时再根据实际情况做出决定。

（二）征求意见

1．完成《信息处理系统语言文字评测规范》（征求意见稿）

2006 年 2 月，课题组完成了《信息处理系统语言文字评测规范》（征求意见稿），并向专家征求意见，编制了《专家意见汇总处理表 1》。

由于自然语言信息处理系统涉及内容广泛，根据项目任务要求，本规范主要提出了语音合成和文语转换、语音识别、机器翻译、语料库等四种自然语言信息处理系统中的语言文字评测规范。通过上述四种系统，规范也给出了信息处理系统语言文字评测的一般原则和方法，其他系统有关评测可以参考。

2. 把征求意见稿分解成《文语转换和语音识别系统语言文字评测规范》《机器翻译系统语言文字评测规范》《语料库系统语言文字评测规范》三个规范草案

2006 年 3 月 3 日，我们把《信息处理系统语言文字评测规范》（征求意见稿）以及《专

家意见汇总处理表 1》提交给教育部语言文字信息管理司，司领导委托国家语言文字委员会语言文字规范（标准）审定委员会语法分委员会（挂靠在北京大中文系）进行审定，分委员会在更加广泛的范围内，第二次征求专家意见，并于 2006 年 6 月 11 日编制了《专家意见汇总处理表 2》。

大多数专家认为，《信息处理系统语言文字评测规范》的范围太大，目前我们的工作难以涵盖规范题目的内容，因此，他们强烈建议把这个规范进行分解，在目前的条件下，先制订《文语转换和语音识别系统语言文字评测规范》《机器翻译系统语言文字评测规范》《语料库系统语言文字评测规范》三个规范，至于其他信息处理系统语言文字评测的问题，今后有条件的时候再制定其他规范。根据专家的意见，课题组对原规范进行了分解，于 2006 年 7 月形成了三个规范草案，准备提交鉴定。

（三）鉴定

2006 年 8 月 20 日，教育部语言文字信息管理司领导委托国家语言文字委员会语言文字规范（标准）审定委员会语法分委员会对于课题组提交的三个规范草案进行鉴定。鉴定会由北京大学计算语言学俞士汶教授担任鉴定组组长，与会专家仔细地审查了课题组提供的各种技术资料和文件，经过认真的讨论，一致通过《文语转换和语音识别系统语言文字评测规范》（草案）《机器翻译系统语言文字评测规范》（草案）《语料库系统语言文字评测规范》（草案）三个软规范。

三　研制原则和方法

（一）原则

规范是人们知识和经验的总结。我们研制的规范必须从国内外自然语言处理系统的评测的实践经验中来，而不能凭空想象。因此，这三个规范研制的原则是尊重自然语言处理系统评测的经验，学习自然语言处理系统评测的经验，在经验的基础上加以总结，升华为语言信息处理系统评测的规范。“从语言信息处理评测的实践中来，到语言信息处理评测的实践中去”，这是我们始终坚持的研制原则。

（二）方法

根据这样的研制原则，我们研制规范的方法，当然也主要是调查研究的方法，我们曾经调查过国内外语言信息处理系统评测的详细情况，写了 7 个调查报告，报告的篇幅超过 10 万字。这 7 个调查报告成为了我们研究规范的最重要的信息资源。

四　研制内容

（一）主要内容

下面我们对于在研制上述三个规范的过程中所采用的一些评测方法的主要内容和科学原理做简要的说明。

1. 用于音节清晰度测试的诊断性押韵测试法（Diagnostic Rhyme Test, DRT）

这种方法根据国家标准《GB/T 13504-1992 汉语清晰度诊断押韵测试（DRT）法》和《GB/T 16532-1996 通信设备清晰度 DRT 法评价用语音材料库》来进行汉语音节清晰度测试。

诊断性押韵测试法用来测试语音编码器输出语音的清晰度，在采用诊断性压韵测试法测试时，每个系统一般使用两张 DRT 音节表，三个音节为一组，每个组为一个文本文件，扩展名为 “.TXT”。测试时，规范发音速度应为 3-4 音节/秒，不应太慢。语音合成系统输出时，三个音节为一组，前加引导语“第 NN 组是”，格式为：

“第 NN 组是 XXX” （NN 为序号，XXX 为测试音节）。

例如，在测试时，如果每个系统使用两张 DRT 音节表，每个表测试 75 个音节，三个音节为一组，75 个音节共 25 组，一共可以测试 150 个音节。

下面是 DRT 音节表的实例：

第一张 DRT 音节表：

第 1 组是选汉因	第 1 4 组是北哲涩
第 2 组是英知绑	第 1 5 组是饶烘鹅
第 3 组是狗说更	第 1 6 组是枕射乏
第 4 组是乡在甲	第 1 7 组是丢绍笨
第 5 组是前也懦	第 1 8 组是化庆卧
第 6 组是摇决门	第 1 9 组是苦敌晒
第 7 组是胡自泡	第 2 0 组是亿二暗
第 8 组是离贴爱	第 2 1 组是律十父
第 9 组是亭入七	第 2 2 组是九边国
第 1 0 组是懒盾见	第 2 3 组是搭追姑
第 1 1 组是脚密信	第 2 4 组是习诈尺
第 1 2 组是宗等穿	第 2 5 组是抹尼上
第 1 3 组是痛词委	

第二张 DRT 音节表：

第 1 组是元十恩	第 6 组是吉蛇收
第 2 组是习凝过	第 7 组是在外由
第 3 组是住树工	第 8 组是董麻汤
第 4 组是红考赤	第 9 组是养哈郑
第 5 组是瑞车乱	第 1 0 组是眼倍搞

第１１组是鱼件门
第１２组是吕志借
第１３组是叫足辽
第１４组是显次米
第１５组是豆晒批
第１６组是印死凶
第１７组是恰代奖
第１８组是茄地一
第１９组是退者心
第２０组是英左黄
第２１组是奴日展
第２２组是更补发
第２３组是纺德善
第２４组是拖运捉
第２５组是必旧跟

编制这样的 DRT 测试音节表的原则是：

a）注意选择那些在语音合成中最有代表性的音节；

b）注意选择那些在语音合成中容易出现错误的音节；

c）注意参照《普通话异读词审音表》（1985 年 12 月修订），选择《普通话异读词审音表》中一些规定必须统读的字来进行测试。如：异读字“胞”有[bao1] [pao1]两个读音，应统读为[bao1]；异读字“呆”有[dai1] [ai2]两个读音，应统读为[dai1]；异读字“酵”有[xiao4][jiao4]两个读音，应统读为[xiao4]，以遵守国家规范。

2．用于音节清晰度测试的改进的压韵测试法（Modified Rhyme Test, MRT）

MRT 是一种改进的压韵测试。DRT 中，每组可测三个汉字的读音，而在 MRT 中，每组只测一个汉字的读音，。这种 MRT 测试法的结果，可以作为 DRT 测试法的参考。

使用 MRT 测试时，每个系统一般使用四个 MRT 表来测试，每个 MRT 表包含若干组，每组为一个文本文件，扩展名为“.TXT”。每个音节嵌入负载句里，前加序号，以正常语速输出。输出格式是：

“N 我读 X 字”

其中，N 是序号，X 代表被测音节，按顺序排列。

例如，在测试时，如果每个系统使用四张 MRT 表，每表 22 个音节，一共可测试 88 个音节。

下面是 MRT 音节表的实例：

第一张 WRT 音节表：

1 我读选字
2 我读英字
3 我读狗字
4 我读乡字
5 我读前字
6 我读摇字
7 我读胡字
8 我读离字
9 我读亭字
１０ 我读懒字
１１ 我读脚字
１２ 我读宗字
１３ 我读痛字
１４ 我读北字
１５ 我读饶字
１６ 我读枕字
１７ 我读丢字
１８ 我读化字
１９ 我读苦字
２０ 我读亿字
２１ 我读律字
２２ 我读九字

第二张 MRT 音节表：

1 我读元字	12 我读吕字
2 我读习字	13 我读叫字
3 我读住字	14 我读显字
4 我读红字	15 我读豆字
5 我读瑞字	16 我读印字
6 我读吉字	17 我读恰字
7 我读在字	18 我读茄字
8 我读董字	19 我读退字
9 我读养字	20 我读英字
10 我读眼字	21 我读奴字
11 我读鱼字	22 我读通字

在编制 MRT 表时，也应当遵循编制 DRT 音节表的原则，选择语音合成中具有代表性的音节和容易出现错误的音节，并注意选择《审音表》中的统读字。

3．用于单词清晰度测试的语义不可预测句(Semantic Unpredicable Sentence, SUS)

单词清晰度测试时，每个系统一般使用两张词表，表中包含若干语义不可预测句，每个语义不可预测句为一个文本文件，扩展名为“.TXT”，格式为：

“1 西瓜 吃 黑色 太阳”

其中，数字表示序号，词间的空格起分词作用，系统输出时，在序号后稍做停顿（300ms），四个词以句子的方式读出；词间不加停顿（空白段不能超过 50ms）。

例如，在测试时，如果每个系统使用两张词表，每表由 25 个语义不可预测句组成，每句包含四个单词，这样，每个表就可包含 100 个单词。一个语义不可预测句为一个文本文件，文件扩展名为“.TXT”，共 25 个文本文件。 例如，

1	隔壁	电报	懂	扣子
2	少有的	服务	蒸	飞机
3	学问	房子	马	锤子
4	粒	安全	三十	人们
5	爷爷	下去	一	碟子
6	天	图书馆	主动	时间
7	外甥	回答	前途	球
8	满意	交代	破	修养
9	老太太	拥护	白	道德
10	紧急	明年	豆	恼怒
11	堆	灾荒	尊敬	商业
12	握手	软	真正	伙食
13	毛线	十八	活	日历
14	这样	勇敢	黄	工资

15	电车	粉条儿	狠	那里
16	担心	领子	癌	开车
17	五月	操场	烧	墨盒
18	神经病	静	紧张	哥哥
19	鲫鱼	簸箕	洗	所谓
20	收音机	证据	归	公里
21	布鞋	罢工	五	唱歌
22	市场	腰	通讯员	四十
23	以内	说话	锅	树皮
24	小时	过去	蘑	不错
25	火车站	窄	溺水	气

4．用于单句清晰度测试的语义可预测的句子

测试单句的清晰度，主要测试单句中语调的清晰性和单句类型的差异性，因此，可以采用语义可预测的句子来进行测试。每个系统一般采用两张句表，句表中包含若干句子，每个句子为一个文本文件，文件扩展名为.TXT。语音合成系统输出时，每句前加序号。格式为：

“15 近几年来上海的国债市场交易活跃”

其中，前面的数字表示序号，系统输出时，在序号后稍做停顿（300ms）。

在测试单句清晰度时，应该注意区分单句的类型（如，陈述句、命令句、感叹句、疑问句），注意测试标点符号对于句子语音清晰度的影响。

例如，在测试时，如果每个系统使用两张句表，每张句表由 25 个句子组成，一共可测试 50 个句子。

下面是句表的实例：

第一张句表：

1 经济起飞象神话似的创造出了奇迹。

2 他成了大学教授，不忘人民养育之恩。

3 在人的一生中，机会可遇而不可求！

4 通过国旗法进行爱国主义思想教育。

5 实验失败了 3 2 次，难度可想而知。

6 就象在家里一样，你可千万不要客气！

7 国家规定不能拿社会赞助去发奖金。

8 购买金融债券儿是个好的投资方向。

9 改革中要抓住机遇需要有时间观念。

10 警察在游戏厅没收赌博机 7 3 台。

11 李瑞环在机场发表了简短书面讲话。

12 积极参加储蓄功在国家，利在人民。

13 亚洲各国政府绝不能放松粮食生产。

14 工厂里培养人才有没有质量标准呢？

15 销路好，这几天要货的电话响个不停。

16 世界各国治理环境都要抓紧进行。

１７　十年来我国民用航空事业发展迅速。

１８　宣传教育法要多样化，不要概念化。

１９　青年人对职业道德教育有哪些想法儿？

２０　陕西省秦兵马俑是人类的伟大创造。

２１　外国重视中国改革开放的大好时期。

２２　只有６．４％的人反对，提案被通过了。

２３　纪念“五四”运动，发扬革命传统。

２４　工人用生命和鲜血保护了国家财产。

２５　纸价大幅上涨，图书市场销售困难。

第二张句表：

１　子弟兵把人民当亲人，对百姓如父母。

２　客厅里，火炉中的木柴烧得越来越旺。

３　殴打病人损害了“白衣天使”的形象。

４　中国名牌商品发展的困难是什么？

５　外语、计算机、会计等是热门儿专业。

６　加强领导，娱乐业要摆脱低俗形象。

７　为什么这几年师范大学毕业生抢手？

８　据报道去年中国经济增长９．４％　。

９　工资打白条儿使教师的生活更加困难。

１０　历史证明没有文化的民族很难自立。

１１　在各种美术作品中他最喜欢国画儿。

１２　引进国外技术的同时要注意国产化。

１３　社会快速发展，最难把握的是机遇。

１４　大家认为听歌剧是高雅的艺术享受。

１５　语言是人类社会的重要交际手段。

１６　发展教育事业中国未来才有希望。

１７　一下子来了１００多位民工，真不少！

１８　经常听广播是他生活中最大的快乐。

１９　人人努力工作，对社会应有所贡献。

２０　幸福、长寿是人类共同的美好愿望。

２１　政府决定克林顿夫妇补交税款１４万。

２２　虽然报纸不断提醒，还有许多人上当。

２３　夜总会有很浓的商业色彩，叫人讨厌。

２４　老师说：请同学们掂掂这钱的分量。

２５　异想天开，一夜之间就想变成个富翁！

单句清晰度也可以采用语义不可预测句子来测试，从而避免听音人根据上文猜测下文和学习效应等弊病，以提高单句清晰度测试的准确性。这样的单句清晰度测试表由若干个语义不可预测的句子组成。这些句子在构成上符合语法，并且尽可能覆盖测试语种的典型语法结构，但在语义上，是独立于上下文的，是不可预测的。

例如：　风平浪静的　苹果　朝着　木樨园　跋涉。

别 让 她 在 青翠 剧烈的 手榴弹 上 输电。

淡漠的 糕点 控制 干燥的 面条。

清闲的 价值 繁殖 清晰的 火柴。

富饶的 混合泳 把 前场 给了 跳水。

你们 递给 旅游区 六 朵 俱乐部。

单句清晰度测试与语音的自然度有密切联系，在测试单句的清晰度时，还应当注意语音的自然度。

5．用于人工评测的两两比较评分法和平均评价分评分法

1）两两比较评分法（Paired Comparison, PC）

以语音合成系统的评测为例。如果参测系统为 n 个，每个系统的语音合成结果都要和其他的 n-1 系统的合成结果对比，这样共有 n(n-1)/2 个对比组合。m 段短文就有 n(n-1)m/2 个两两对比组合（每个组合由 A 和 B 组成，A、B 表示同一段短文的不同的语音合成系统的合成版本）。听音人按照韵律、流畅性，拟人性等因素对合成结果采用五级评分制（+2，+1，0，-1，-2）评分：

- 如果 A 比 B 好，则 A 得+2 分，B 得-2 分；
- 如果 A 比 B 稍微好，则 A 得+1 分，B 得-1 分；
- 如果 A 和 B 不相上下，两者各得 0 分。

为保证评测的公平性和尽可能地削弱学习效应，每个参测单位合成的语音在组合中先后播放的机会均等。如参测单位 A 和 B 比较，在随机选取的 m/2 篇短文中，按顺序 AB 播放；剩下的 m/2 篇短文按顺序 BA 播放。

2）平均评价分评分法（Mean Opinion Score, MOS）

以语音合成系统的评测为例。这种评分法对于听音人要求较高，最好邀请有经验的听音专家来参与评分。听音专家根据语音合成结果的输出与自然语言接近程度的总体印象，从拟人性、连贯性、韵律感等方面，用优、良、中、差、劣五级记分来评价，言语自然度评价的 MOS 分级为：

1 －－ 劣，不能接受

2 －－ 差，不愿接受

3 －－ 中，可以接受

4 －－ 良，愿意接受

5 －－ 优，很自然

参测系统各段得分平均作为其自然度得分。

在得出了 PC 法的评分之后，再请听音专家对用 PC 法评分时平均总分第一名的系统进行 MOS 评分，作为总体评价的参考。在进行 MOS 评分时，各系统短文应打乱顺序播放。

评分时，以 PC 法的评分为主，以 MOS 法的评分为辅。

在用 MOS 法评分时，如果听音专家不清楚五级记分的具体表现就可能出现大的方差。可以把好的语音和坏的语音例子让听音专家先听一下，然后再开始测试打分。在同样的测试条件下，在不同国家用本土语言，听音专家不容易在等级定位上取得相互一致。因此，MOS 需要进行调整之后，才可

以得到可靠的品质指标。

（二）适用范围

我们制定的三个规范提出了文语转换和语音识别系统、机器翻译系统和语料库系统评测的一般原则和具体方法。

这三个规范可供文语转换和语音识别系统评测、机器翻译系统评测和语料库系统评测以及有关管理工作参考采用。

（三）主要争议问题的处理

我们对于在国内外学术界主要争议的词错误率（word error rate）的计算方法进行了认真的分析和适当的处理。

国际上语音识别系统的标准评测指标是词错误率。词错误率的计算是根据语音识别系统返回的单词串（通常叫做假定单词串[hypothesized word string]）与标准答案或参照转写（reference transcription）中的正确单词串的差别的大小来进行的。给定一个正确的转写，计算词错误率的第一步是计算在假定单词串和正确单词串之间的单词的最小编辑距离（叫做 Leventhtain 距离）。这个计算的结果将是在正确单词串和假定单词串之间映射时单词替代、单词插入和单词脱落的最小数目。这样，词错误率可以定义如下（注意，由于等式中包含插入，错误率可能会大于 100%）：

$$\text{词错误率} = 100\times\frac{\text{单词插入数}+\text{单词替代数}+\text{单词脱落数}}{\text{正确转写的单词总数}}$$

下面是来自 CALLHOME 语料库中一个例子，参照单词串和假定单词串上下对齐，同时还给出了用于计算词错误率的计数：

参照单词串：	i	***	**	UM	the	PHONE	IS	i	LEFT	THE	portable
假定单词串：	i	GOT	IT	TO	the	*****	FULLEST	i	LOVE	TO	portable
评估：		I	I	S		D	S		S	S	

参照单词串：	****	PHONE	UPSTAIRS	last	night	so	the	battery	ran	out
假定单词串：	FORM	OF	STORES	last	night	so	the	battery	ran	out
评估：	I	S	S							

这个语段有 6 个替代（S），3 个插入（I），1 个脱漏（D）：

$$\text{词错误率} = 100\times\frac{6+3+1}{18} = 56\%$$

最近，国外使用美国国立标准和技术研究所（NIST）的广播新闻语料 Hub4 测试集来测试英语自然语音，最新的英语语音识别系统的词错误率达到 20%（Chen 等，1999），如果用 Hub5 测试集再加上 Switchboard，Switchboard-II 和 CALLHOME 语料库来测试自然语音，词错误率为 40%（Hain 等，1999）。

仿照“词错误率”的这个定义，在汉语语音识别评测中，可提出相应的评测指标。汉语语音识别系统的标准评测指标是汉字错误率（character error rate）。

汉字错误率的计算是根据语音识别系统返回的汉字串（通常叫做假定汉字串[hypothesized

character string]，也就是识别结果）与正确转写或参照转写（reference transcription，也就是标准答案）的差别的大小来进行的。给定一个正确的转写，计算词错误率的第一步是计算在假定汉字串和正确汉字串之间的汉字的最小编辑距离。这个计算的结果将是在正确汉字串和假定汉字串之间映射时汉字替代、汉字插入和汉字脱落的最小数目。这样，汉字错误率可以定义如下：

$$\text{汉字错误率} = 100 \times \frac{\text{汉字插入数}+\text{汉字替代数}+\text{汉字脱落数}}{\text{正确转写的汉字总数}}$$

由于等式中包含插入汉字数，错误率可能会大于 100%。

汉语的例子：

标准答案：	历时	三	天	三	夜	*	顾	不	上	休息
识别结果：	历*	三	田	伞	也	勾	顾	布	尚	休息
评估：	D		S	S	S	I		S	S	

这个语段有 5 个替代（S），1 个插入（I），1 个脱漏（D）：

$$\text{汉字错误率} = 100 \times \frac{5+1+1}{11} = 63.63\%$$

仿照“汉字错误率”的这个定义，在语音识别评测中，可以提出相应的评测指标如下。

采用动态规划算法求Levenshtein距离的全局最小值，以得到“正确文字数”、“替换文字数”、“插入文字数”、“删除文字数”四项参数。（“文字”在汉语中指字，在外语中指单词，下同）。并且规定：

原文答案文字数 ＝ 正确 ＋ 替换 ＋ 删除文字数；

输出结果文字数 ＝ 正确 ＋ 替换 ＋ 插入文字数；

定义：

错误文字数 ＝ 替换 ＋ 插入 ＋ 删除文字数；

语音识别的评价主要用以下五个指标来衡量：

- 文字错误率 ＝ 错误文字数/原文答案文字数×100％；
- 插入文字错误率 ＝ 插入错误文字数/原文答案文字数×100％；
- 删除文字错误率 ＝ 删除错误文字数/原文答案文字数×100％；
- 替换文字错误率 ＝ 替换错误文字数/原文答案文字数×100％；
- 句子正识率 ＝ 正识（一字不差）句子数/句子总数×100％。

此外还用其他两个指标来辅助衡量整个系统的性能：

- 完成时间 ＝ 系统完成全部任务所用的时间；
- 系统故障数 ＝ 在评测过程中测试系统出现故障的次数。

我们在规范第一稿的制订过程中，认真研究了国外关于“词错误率”的计算方法，也研究了“汉字错误率”的计算方法，并结合我国评测工作的具体情况，对这些指标的计算方式做了一些修正，主要的修正之处是，在第一稿中使用“正确＋替换＋插入＋删除汉字数目”来替换公式中的“原文答案文字数”，其原因在于，“原文答案文字数”中没有包括“删除汉字数目”（也就是在“原文答案”中有，而在输出结果中没有的文字数），而“删除汉字数目”对于评价错误率是有意义的。第一稿中的计算公式如下：

语音文本转换的评测指标：

• 汉字正识率 ＝ 正确汉字数/原文汉字总数×100％；
• 插入汉字错误率 ＝ 插入错误汉字数目/（正确＋替换＋插入＋删除汉字数目）×100％；
• 删除汉字错误率 ＝ 删除错误汉字数目/（正确＋替换＋插入＋删除汉字数目）×100％；
• 替换汉字错误率 ＝ 替换错误汉字数目/（正确＋替换＋插入＋删除汉字数目）×100％；
• 句子正识率 ＝ 正识（一字不差）句子数/句子总数×100％；
• 完成时间 = 系统完成全部任务所用的时间；
• 系统故障数 = 在评测过程中测试系统出现故障的次数。

音节识别的评测指标：

• 音节正识率 ＝ 正确音节数/原文音节总数×100％；
• 插入音节错误率 ＝ 插入错误音节数目/（正确＋替换＋插入＋删除音节数目）×100％；
• 删除音节错误率 ＝ 删除错误音节数目/（正确＋替换＋插入＋删除音节数目）×100％；
• 替换音节错误率 ＝ 替换错误音节数目/（正确＋替换＋插入＋删除音节数目）×100％；
• 完成时间 = 系统完成全部任务所用的时间；
• 系统故障数 = 在评测过程中测试系统出现故障的次数。

在书面征求意见时，许多专家认为，评测指标应当与国际接轨，最好不要另稿一套，因此，我们在新的修订稿中，根据专家意见，仍然按国际通用的方式进行计算，使用“原文答案文字数”来替换“正确＋替换＋插入＋删除音节数目”。

五　其他

（一）与相关标准的关系

由信息产业部中文语音交互技术标准工作组和信息产业部标准化研究所合作提出的《中文语音合成系统通用技术规范》、《中文语音识别系统通用技术规范》已经处于国家标准的报批中。我们在研制本规范时注意到了他们的工作，在有关语言文字的评测规范方面，力求与他们保持一致。

（二）废止相关标准的建议

无。

（三）使用建议

由于我们的这三个规范都是软规范，它们只有推荐性，没有强制性，目前还是规范草案，建议用户在使用过程中不断发现问题，并且给我们提出来，供今后进一步修时参考。

六　相关资料

（一）参考文献

1．ISO11179-3 信息技术 数据元的规范与标准化　第 3 部分：数据元的基本属性
2．GB/T 12200.1　　汉语信息处理词汇 01 部分：基本术语
3．GB/T 13725　　信息处理用现代汉语分词规范
4. GB/T 16403-1996　　声学 测听方法 纯音气导和骨听阈基本测听法，eqv ISO 8253-1:1989
5. GB/T 17696-1999　　声学 测听方法 第 3 部分：语言测听，eqv ISO 8253-3:1996
6. GB/T 16404.1-1996　声学 声强法测定噪声源的声功率级 第 1 部分：离散点上的测量，eqv ISO 9614-1:1993
7. GB/T 16404.2-1999　声学 声强法测定噪声源的声功率级 第 2 部分：扫描测量（Acoustics-Audiometric test methods-Part 3: Speech audiometry），eqv ISO 9614-2:1996
8. GB/T 13504-1992　　汉语清晰度诊断押韵测试（DRT）法（Diagnostic rhyme test（DRT） method

of Chinese articulation）
9. GB/T 16532-1996　　通信设备清晰度 DRT 法评价用语音材料库（Speech material library used DRT for articulation evaluation of communication equipments），idt IEC 728-1:1986
10. GB/T 17147-1997　　声音广播中音频噪声电平的测量，eqv ITU-R 468-4:1990
11. GB/T 17576-1998 CD　　数字音频系统，idt IEC 908:1987
12. GB/T 14476-1993　　客观评价厅堂语言可懂度的 RASTI 法，neq IEC 268-16
13. GB/T 14919-1994　　数字声音信号源编码技术规范（The specifications for digital audio source coding），eqv CCIR 646
14. GB/T 16463-1996　　广播节目声音质量主观评价方法和技术指标要求
15. SJ 20771-2000　　军用通讯系统音质 MOS 评价法
16. GB3259－92　　中文书刊名称汉语拼音拼写法
17. GB/T 15834－1995 标点符号用法
18. GB/T 15835－1995 出版物上数字用法的规定
19. GB/T 12200.1　　汉语信息处理词汇 01 部分：基本术语
20. GB/T 13725　　信息处理用现代汉语分词规范
21. GB3259－92　　中文书刊名称汉语拼音拼写法
22. GB/T 16785-1997　　术语工作　概念与术语的协调
23. GB/T 16786-1997　　术语工作　计算机应用　数据类目
24. GB/T 17532-1998　　术语工作　计算机应用　词汇
25. GB/T 15834－1995 标点符号用法
26. GB/T 15835－1995 出版物上数字用法的规定
27. GB/T 16159－1996 汉语拼音正词法基本规则
28. 第一批异形词整理表
29. 第一批异体字整理表
30. 部分计量单位名称统一用字表
31. 中国人名汉语拼音字母拼写法
32. 中国地名汉语拼音字母拼写规则
33. 普通话异读词审音表
34. J. White, Approaches to black-box MT evaluation, Proceedings of MT Summit V, Luxembourg, 1995.
35. Battelle, The evaluation and system analysis of the SYSTRAN machine translation system, RADC-TR-76-399 Final technical report, Battelle Columbus Laboratories, Rome Air Development Center, Air Force System Command, Griffs Air Force Base, New York, 1977.
36. K. Falkedal, Evaluation method for machine translation system: an historical overview and critical account, Report ISSCO, Université de Genève, 1991.
37. J. Fiscus, G. Doddington, J. Garofolo, A. Martin, NIST 1998 topic detection and tracking evaluation, Proceedings of the DARPA Broadcast News Workshop, San Francisco, 1999.

（二）参加工作的单位和人员

研制单位：教育部语言文字应用研究所
研制负责人：　冯志伟
主要研制人：　肖航，富丽
参与研制人：　章云帆
提供帮助的专家：王仁华
鉴定组组长：　俞士汶
鉴定组成员：　常宝宝，詹卫东，刘群，李爱军，林守勋，沈玉兰，靳光谨

2008 年 10 月 1 日