中华人民共和国国家标准

汉语信息处理词汇 01部分:基本术语

GB 12200.1—90

Chinese information processing—Vocabulary
Part 01:Fundamental terms

本词汇涉及到汉语信息处理的各个主要方面,其中包括基本术语、汉语和汉字、汉字编码、汉字识别、汉语语音处理、汉语理解、机器翻译、汉语信息处理设备、汉语信息处理系统软件、汉语信息处理技术应用及其他等约11个部分。在学科方面,本词汇具有相对的独立性和系统性。

1 主题内容与适用范围

1.1 主题内容

本标准规定了最重要的或最基本的汉语信息处理术语,它们是其他各部分的基础。

1.2 适用范围

本标准适用于有关汉语信息处理领域的科研、设计、生产、使用、维护、管理、教学和出版等方面。

2 引用标准

GB 2312 信息交换用汉字编码字符集 基本集

GB 5271 数据处理词汇

3 遵循的原则和规则

3.1 词条

词条是本标准为使用者提供的便于查检和参阅的基本单元。

3.2 词条的组成

本标准的词条一般由下述几部分构成:

a. 索引号(不同语种文本都是一致的);

b. 术语;

c. 术语的缩写;

d. 允许用的同义术语;

e. 术语的英译名;

f. 术语的英文缩写;

g. 术语的定义;

h. 以"例:"开头的一个或几个示例;

i. 以"注:"开头的一个或几个注释(用以说明术语应用的特殊情况);

j. 图、图表或表格。

3.3 多义术语

当一个术语有几个不同的意义时,分别在不同的词条中给予定义,以便于译成其他语种。

国家技术监督局1990-02-01批准 1990-08-01实施

3.4 缩写

有些术语具有常用的缩写，但在定义、示例及注释中，不采用这种缩写。

3.5 符号的用法

3.5.1 圆括号的用法

有些术语使用时，在不引起误解的情况下，可以省略掉其中一部分，可省略的部分为黑体字，并用圆括号括起。在定义、示例和注释中，只用完整的术语。

有些术语后圆括号内的非黑体字，不是术语的组成部分，而是用来说明该术语的使用须知和特殊应用形式或语法形式的。

3.5.2 方括号的用法

当几个术语使用同一个定义格式(个别词不相同)时，可将它们合并在一个词条中。个别不相同的词放在方括号中，表示可以替换方括号前面的词。方括号及其中的词在术语及定义中出现的顺序必须一致。

3.5.3 黑体字与星号

术语在定义、示例和注释中用黑体字印刷时，表示该术语已在本词汇的其它词条中给过定义，并且只有它在另一词条中首次出现时才印成黑体字。

如果有两个已分别在不同的词条中给过定义的术语连在一起使用时，则用星号"*"将这两个术语隔开。

3.6 英译名

术语所对应的英文采用美国习用的拼法。

3.7 索引

本标准附有汉语索引和英文索引。索引包括本部分的全部术语。

根据汉语索引或英文索引，可查出术语正文的索引号。

4 术语和定义

4.1 基本术语

4.1.1 一般术语

4.1.1.1 语言信息处理 language information processing

用计算机对**自然语言**的音、形、义等信息进行处理。即对字、**词**、句、篇章的输入、输出、识别、分析、理解、生成等的操作与加工。

4.1.1.2 汉语信息处理 Chinese information processing

用计算机对**汉语**的音、形、义等信息进行处理，有时又称中文信息处理。

4.1.1.3 汉字信息处理 Chinese character information processing

用计算机对**汉字**表示的信息进行的操作和加工，如汉字的输入、输出、识别等。

4.1.1.4 汉字输入 Chinese character input

利用**汉字**的形、音或相关信息通过各种方式，把汉字输入到计算机中去的过程。

4.1.1.5 汉字输出 Chinese character output

将计算机内以数据形式表示的**汉字**在显示终端、印字机等设备输出的过程。

4.1.1.6 多文种信息处理 multilingual information processing

在两种或两种以上**语言**·**文字**字符集编码体系基础上，实现对多文种信息的兼容处理。

4.1.1.7 民族语言支撑能力 National language support NLS(缩写)

使计算机具备能够处理民族语言的开发能力。

例：中文化，汉字化。

4.1.2 语言文字

4.1.2.1 语言 language
为了传递**信息**而使用的一组**字符**、约定和规则。
注：同 GB 5271.7 的 07.02.01 条。

4.1.2.2 自然语言 natural language
一种**语言**，其规则是根据当前流行的用法而不是用明确的形式规定的。
注：同 GB 5271.7 的 07.02.03 条。

4.1.2.3 人工语言 artificial language
一种**语言**，其规则在使用前已明确地规定了。
注：同 GB 5271.7 的 07.02.03 条。

4.1.2.4 受限语言 restricted language
在**词汇**、**句法**、**语义**及**语用**等方面受到人为限制的**自然语言**的真子集。

4.1.2.5 语言模型 linguistic model
对**自然语言**的数学描述。分为生成模型、分析模型和识别模型三种。

4.1.2.6 语音 speech sound
人类发出的能表达一定意义的声音。

4.1.2.7 文字 script
人类记录和传达**语言**的书写符号系统。

4.1.2.8 词 word
最小的能独立运用的**语言**单位。
例：大、国家、奥林匹克。

4.1.2.9 词汇 vocabulary
一种**语言**中所有的**词**与固定词组的集合。

4.1.2.10 语法 grammar
语言的结构规则。**自然语言**的语法具有一定的民族特点和相当的稳定性。

4.1.2.11 句法 syntax
词或词组之间的组合规则。

4.1.2.12 语义 semantics
词或词组与它们的含义之间的关系。

4.1.2.13 语用 pragmatics
词或词组与它们的解释和使用之间的关系。

4.1.2.14 文本 text
语言的符号串，**文字**信息处理的对象。

4.1.2.15 语言资料库 corpus
文本的有序集合。各种分类、检索、综合、比较的基础。

4.1.2.16 语言知识库 language knowledge base
计算机所存储的**语言**知识的集合。它是计算机从**语音**、**文字**、**词汇**、**句法**、**语义**、**语用**等角度对语言进行信息处理的基础。

4.1.2.17 计算语言学 computational linguistics
语言学的一个分支学科。它应用计算机技术来研究和处理**语言**、**文字**，内容包括：字频和词频统计、**语音**的识别与合成、**机器词典**的编纂、**机器翻译**、自然语言理解、计算机的自然语言接口等。

4.1.3 汉语和汉字

4.1.3.1 汉语 Chinese

汉族的**语言**。中国境内主要的通用语言，也是国际通用语言之一。属汉藏语系。

4.1.3.2 中文 Chinese

特指**汉语**。

4.1.3.3 现代汉语 contemporary Chinese language

现代汉民族语言。包括它的主要地域分支：北方话、吴语、湘语、赣语、粤语、客家话、闽语等。它的规范化语言是**普通话**。

4.1.3.4 普通话 Putonghua

现代汉民族共同语。它是规范化的**现代汉语**，以北京语音为标准音，以北方话为基础方言，以典范的现代白话文著作为语法规范。

4.1.3.5 汉语拼音(方案) scheme of the Chinese phonetic alphabet, Pinyin

给**汉字**注音和拼写**汉语**·**普通话**·**语音**的方案。方案采用 26 个拉丁字母，有声母表和韵母表及拼写规则，对声调符号和隔音符号的标记也有规定。

4.1.3.6 汉字 Chinese character, Hanzi

记录**汉语**的书写符号系统。汉字也被其他一些国家或民族用作为书写符号。

4.1.3.7 现代通用汉字 current commonly-used Chinese character

现代通行的记录**现代汉语**的书写符号系统。

例：(1)GB 2312。

(2)《现代汉语通用字表》。

4.1.3.8 汉字属性 attribute of Chinese characters

汉字本身所具有的音、形、义三方面的特征及附加的有关特征。

例：笔画、笔顺、部首、部件、**汉语拼音方案**、四角号码等。

4.1.3.9 汉字属性字典 Chinese character attribute dictionary

包括**汉字**部首、**汉语拼音方案**、笔画数、笔顺、使用频度、组词能力、文字结构、标准部件、标准字形点阵码等属性及其电报码等相关信息的数据库或字典。

4.1.3.10 简化字 simplified Chinese character

采用同音代替、改换声旁、草书楷化、偏旁简化类推等方法制定的一批笔画较少的**汉字**。这些字取代了对应的笔画较多的汉字作为通行的正体。特指 1986 年重新公布的《简化字总表》，共 2 235 字。

例：后[後]，亿[億]，发[發]，说[說]，难[難]。

4.1.3.11 繁体字 unsimplified Chinese character

被**简化字**代替的笔画较多的**汉字**。

例：专[專]，圣[聖]，宁[寧]，对[對]，机[機]。

4.1.3.12 异体字 variant Chinese character

汉字通常写法之外的一种音同、义同，只是字形笔画或结构不同的字。

例：升[昇、陞]，迹[跡、蹟]，泪[淚]。

4.1.3.13 分词单位 word segmentation unit

汉语信息处理使用的、具有确定的**语义**和(或)**语法**功能的基本单位。

4.1.3.14 汉语分词 Chinese word segmenting

从工程观点出发，按照特定的规范，对**汉语**按**分词单位**进行划分的过程。

4.1.4 汉字编码

4.1.4.1 汉字[汉语词语]编码 Chinese character[Chinese word and phrase]coding

按照一定的规则，对指定的**汉字**[汉语词语]集内的元素编制相应的代码。

4.1.4.2 汉字编码字符集 Chinese character coded character set

按一定的规则确定的包含**汉字**及有关基本图形字符的有序集合，并规定该集合中的字符与编码表示之间一一对应的关系。

例：GB 2312。

4.1.4.3 汉字编码方案 Chinese character coding scheme

汉字集元素映射到其他字符集元素的一组完整规则。

4.1.4.4 汉字编码(键盘)输入方法 Chinese character coding (keyboard) input method

运用某种编码方案、键盘设备及计算机资源，由操作者向计算机输入**汉字**的方法。

4.1.4.5 汉字编码输入方法评估 evaluation of Chinese character coding input method

按照约定的或法定的规则和步骤，对**汉字编码(键盘)输入方法**的素质和特性进行定量的测试和定性的评价等。

4.1.4.6 汉字(信息)交换码 Chinese character code for information interchange

汉字信息处理系统之间或者信息处理系统与通信系统之间进行汉字信息交换的代码。

4.1.4.7 汉字内部码 Chinese character internal code

汉字在信息处理系统内部最基本的表达形式，供存储、处理、传输汉字用。

注：它与**汉字信息交换码**有一定的对应关系，通常借助于某种特定标识信息来表明它与单字节字符的区别。

4.1.4.8 汉字控制功能码 Chinese character control function code

说明**汉字**数据的传送控制、格式处理、汉字扩充及设备控制等的代码。

4.1.4.9 汉字字形码 Chinese character font code

表达**汉字**字形的字模数据，通常用点阵、矢量函数等方式表示。

4.1.4.10 汉字点阵字形 Chinese character dot matrix font

计算机中以点阵形式表示规范化**汉字**字形的一种形式。

4.1.5 语音和文字自动处理

4.1.5.1 汉字识别 Chinese character recognition

利用计算机抽取**汉字**字形特征，实现对汉字的自动输入。可分为联机手写体汉字识别、印刷体汉字识别和手写体汉字识别。

4.1.5.2 汉语语音识别 Chinese speech recognition

利用**语音**分析技术，抽取语音特征，实现对**汉语**语音的自动识别。可分为特定人和非特定人两种。

4.1.5.3 汉语语音分析 Chinese speech analysis

将**汉语**语音模拟信号转换为语音数字信号，抽取汉语语音的特征，建立汉语语音模型的过程。

4.1.5.4 汉语语音合成 Chinese speech synthesis

利用**汉语语音信息库**和**语音**的合成系统，合成出所需**汉字**、单词、短语或整句的汉语语音流。

4.1.5.5 汉语语音信息库 Chinese speech information library

利用**语音**分析、压缩技术，根据**汉语**语音特征模型(声母、韵母、声调、音节、语调、语气)建立的汉语语音数据、参数、特征数据库。

4.1.5.6 汉语语音数字信号处理 Chinese speech digital signal processing

利用**语音**采样、分析、存储、合成、传输等技术，实现对**汉语**语音的识别、录放、合成和通信功能。

4.1.5.7 汉语语音信息处理 Chinese speech information processing

利用计算机系统、**汉语**·**语音**的编码技术和**汉语语音数字信号处理**技术，实现汉语语音输入、输出、理解、翻译、语音和**文字**相互转换以及语音信息通信等信息处理功能。

4.1.6 汉语理解和机器翻译

4.1.6.1 汉语理解 Chinese language understanding

计算机基于语言知识和背景知识对**汉语**进行的分析、判断和推理。

4.1.6.2 机器翻译 machine translation MT(缩写)

用计算机将一种**自然语言**(**源语言**)转换成另一种自然语言(**目标语言**)的过程。

4.1.6.3 机器词典 machine dictionary

以人用词典为基础,通过对词法、**句法**、**语义**等信息的规范化和形式化描述,做成的存储在计算机中的词典。

4.1.6.4 源语言 source language

在**机器翻译**中,被翻译的**语言**。

4.1.6.5 目标语言 target language

在**机器翻译**中,所译成的**语言**。

4.1.6.6 句法语义分析 parsing

用计算机在**句法**和**语法**层次上对句子进行分析,即主要通过**语言**中各种成分的句法功能和**语义**关系来描述句子的结构与层次。

4.1.6.7 汉语分析 Chinese analysis

将输入计算机的**汉语**句子或篇章,利用给定的分析方法,确定每个成分的词法、**句法**和**语义**等信息,并将其转换成便于计算机进一步处理的中间表示。

4.1.6.8 汉语生成 Chinese generation

从计算机中的某种**语言**信息的中间表示出发,通过必要的**语法**和**语义**知识转换生成**汉语**的句子或篇章。

4.1.7 汉语信息处理设备和软件

4.1.7.1 汉语语词处理机 Chinese word processor

一种专用的**汉字信息处理**系统,用于**汉语**文稿的输入、编辑、存储、印刷及传送。

4.1.7.2 多文种语词处理机 multilingual word processor

一种能处理两种及两种以上**语言**·**文字**信息的具有键击输入、显示、校改、文本编辑、印刷输出等功能的设备。该设备通常也具有简单的文档管理功能。

例:中英文电子打字机。

4.1.7.3 汉字印字机 Chinese character printer

能实现中、英文输出的印字设备。通常分为击打式和非击打式两种,一般都配有中、英文字形库。

4.1.7.4 汉字终端 Chinese character terminal

能完成**汉字**输入、输出功能的计算机终端,通常分为简易型、通用型和智能型。

4.1.7.5 汉字输入键盘 Chinese character input keyboard

便于输入**汉字**的专用键盘。它可分为整字型、部件(字根)型等类型。它在键位布局和结构设计等方面有别于传统的西文键盘。

4.1.7.6 汉卡 Chinese character card

将**汉字编码输入方法**的码表和有关程序及**汉字**的字模数据固化在 ROM 器件中的一种逻辑电路插件。

4.1.7.7 汉字字形库 Chinese character font library

建立在计算机存储媒体上的**汉字**的字模数据集合。

4.1.7.8 汉语词语库 Chinese word and phrase library

建立在计算机存储媒体上**汉语**的**词**和短语的集合,该集合可按词语关系的结构作有序的排列,可以按收词多少、词语性质、功能、结构等分类。

4.1.7.9 汉字公用程序 Chinese character utility programs

支援计算机系统在**汉字**方式下实际运行的辅助程序。

例：汉字造字、排序、编辑及打印等程序。

4.1.7.10 多文种信息处理系统 multilingual information processing system

能处理由多种**语言**·**文字**所表述的信息的系统。它们可分为两类：一类是在原有单文种系统基础上扩充而成的，在这类系统中新纳入的文种应用的范围往往受到限制；一类是以能容纳多种文字字符的大字符集为基础设计的系统，可不受上述限制。

4.1.8 汉语信息处理技术应用

4.1.8.1 通用型汉字信息处理系统 general-purpose Chinese character information processing system

适用于各种数据处理和**汉字信息处理**的计算机系统。其特点是通用性强，**汉字**输入输出手段多，操作方便。

4.1.8.2 电子出版系统 electronic publishing system EPS(缩写)

利用以计算机为主的电子技术代替传统的人工编辑、铅字排版的自动化印刷出版系统。

4.1.8.3 中文情报检索系统 Chinese information retrieval system

对中文情报进行收集、存储、检索、加工、编辑和分发传送的信息处理系统。

4.1.8.4 汉语计算机辅助教学系统 Chinese computer-aided instruction system

通过教员或学员与计算机之间的交互活动，辅助编辑教材，选择适于学员个人的学习程序和课程内容，达到教学**汉语**目的的一种信息处理系统。