简繁大师中级教程——简繁转换实战攻略

您可在我们网站上找到本教程的文本稿，有时间可以细看。本教程将着眼于实际操作的演示。

http://www.speedy7.com/cn/stguru/video/

【教程简介】

《简繁大师》可用于对各种编码和格式的文本及文件进行简繁转换，本教程前面对相关编码知识及简繁大师所提供的“专业品质”简繁转换服务进行了介绍，后面则介绍了如何进行各种类型的转换。除了转换功能外，《简繁大师》也是一个很好用的中文文本编辑器，它还有多种词汇管理和语言相关功能，有兴趣的可以通过运行该软件加以了解。

具体目录如下。视频较长，如果您只关心部分内容，可以拖动播放滑动条，快速移动到您感兴趣的部分。

【背景知识：文本/文件的编码】

简繁大师支持在 9 种语言和编码的任意组合之间进行两两互转：

简体中文（GBK）
简体中文（UTF-8）
简体中文（Unicode）
简体中文（Unicode BE）

繁体中文（GBK）
繁体中文（Big5）
繁体中文（UTF-8）
繁体中文（Unicode）
繁体中文（Unicode BE）

GBK 和 Big5

与简繁转换相关的编码主要有 5 种，其中 2 种是单字节编码集 GBK 和 Big5。在这 2 种编码中，一个中文字由两个值在 1-255 之间的单字节字符相连而成（比如“编码”的“编”字，在 GBK 中实际上由值分别为 B1 和 E0 （按 16 进制）或 177 和 224 （按 10 进制）的两个半角字符组成），电脑操作系统又把这两个单字节字符组合在一起，显示为一个汉字字符。

GBK 是由中国大陆制订的，共 20,000 多字，包含了平时我们能碰到的基本所有简体字和繁体字，如与简体字“编”对应的繁体字“編”也在 GBK 编码中。GBK 的前身是 GB2312 编码，GB2312 编码集也是中国大陆制订的，共 6,000 多字，时间较早，只含简体字，由于收编字数较少，现在实际已经被淘汰了。由于 GBK 编码也包含了繁体字，所以我们既可以用它显示简体字，也可以用它显示繁体字。GB2312 编码虽然实际已经淘汰了，但有时我们注意到有些网页文件里标记的编码标记是 GB2312，这些我们平时看到的“GB2312”只是延用了历史上的名称，现实中出现的 GB2312 基本上实际都是指 GBK，不再是老的 GB2312 编码本身。

Big5 是台湾制订的编码集，共 15,000 多字，只含繁体字，所以我们不能用 Big5 编码显示只在简体字符集中出现的简体字。台湾现在逐渐把繁体字改称为“正体字”。

在支持 GBK 或“GB2312”字符的电脑环境中，我们不能直接阅读用 Big5 编码写的字。比如在 Big5 下的繁体字“軟體編碼”在简体中文操作系统中显示为“砐砰緙緀”。如果简体用户碰到港台繁体用户发来的文件或信中有“乱码”，可以试着把这些“乱码”贴入简繁大师的编辑区，并应用编码“繁体中文（Big5）”，一般即可正确显示。

UTF-8、Unicode、Unicode BE

另 3 种编码分别是宽字节编码 Unicode 和它的两种变种 UTF-8 和 Unicode BE。一个 Unicode 字符由一个值在 1-65,535 之间的宽字符组成，比如“编码”的“编”字，在 Unicode

中对应的值是 32534（按 10 进制）或 7F16（按 16 进制）。Unicode 和 GBK 一样，也包含了所有简体字和繁体字（它的中文部分实际上是参照 GBK 建立的），此外，Unicode 还包含了所有其它主流自然语言（如日语、韩语、俄语、法语等）的字符。Unicode 有时也被称之为“统一码”，不过现在主流的写法就是英文形式的“Unicode”。

由于很多操作系统或存贮显示环境并不支持宽字节，所以人们也在 Unicode 基础上编制了 UTF-8 编码，一个 Unicode 字符可以通过某种对应关系一一对应到一组 1-3 个单字节（指值在 1-255 之间）长的 UTF-8 值，由于 UTF-8 兼容单字节电脑环境，所以它现在比 Unicode 本身应用更广泛，已经成为目前最受行业重视的标准编码，基本所有主流文档格式（如 Word、Excel、PowerPoint）都可以导出为 XML 格式文件，而 XML 文件的编码就是 UTF-8 编码。“軟體編碼”转换成 UTF-8 后，看起来的样子是“杌熼玲绶丨12.”。如果你碰到 UTF-8 编码的文字，想看明白它讲什么，可以复制到简繁大师中进行“编辑区转换”-“简体中文（UTF-8）->简体中文（GBK）”，就可以看明白了。由于 GBK 字符集包含了繁体字，所以即使内容是繁体，也可以正确转换，转出来仍是正确的繁体。

Unicode BE 是将 Unicode 编码的前后两部分对换个位置（比如“编”字的 Unicode 编码是 7F16（按 16 进制），在 Unicode BE 中则是 167F），这是某些电脑环境要求的。在 Windows 中 Unicode BE 出镜率不高。

【背景知识：专业品质】

与其它简繁转换引擎不同，简繁大师采用了我们多年深入研发而成的专业品质简繁转换引擎。根据我们的调查，对于大多数商务类、技术类和 IT 类稿件，只要合理选用转换包（普通源稿推荐使用“默认”包、技术类源稿推荐使用“科技”包（即“Science/Tech”包）、IT 行业源稿推荐采用“IT”包），一次性转换即可达到目标语大众读者可以接受的商业品质。

简繁大师的简繁转换品质目前处于行业领先水平。

【文本的简繁转换】

如果您只需要转换一段文本，可以用“编辑区转换码”或“剪贴板转码”。

编辑区转码

如果您需要转换的内容是纯粹的文字，且不包含特殊字符，则编辑区转码就足够了。
编辑区转码的优势是所见即所得。缺点是不支持一些只在 Unicode 编码中才有的特殊字符。
如果您可以确定需要转换的文字的编码是 GBK、Big5 或 UTF-8，推荐您使用编辑区转码。

剪贴板转码

如果您需要转换的内容包含一些特殊字符，且来源编码是 Unicode 或 Unicode BE（如从一些内容中包括特殊字符的网页或其它 Unicode 文件中复制过来的），则建议使用剪贴板“简体中文（Unicode）”与“繁体中文（Unicode）”互转。

【对纯文本文件、网页、目录、网站进行简繁转换】

纯文本文件和网页文件的转换

1）纯文本文件

平时我们需要转换的大部分文本文件的编码是 GBK、Big5、UTF-8（XML 即是 UTF-8）和 Unicode。简繁大师的“文件、网页、目录、网站转码”功能支持对 GBK、Big5、UTF-8、Unicode 和 Unicode BE 进行两两互转。

2）网页文件（如 HTML、ASP 等）

您可以用与转换纯文本相同的方式转换网页文件。简繁大师会自动对网页文件内的语言属性（如 GB2312（即 GBK）、Big5、UTF-8 等）进行转换。最近几年的网页文件以 UTF-8 最多，以前的以 GBK、Big5 为主，另外也有 Unicode 格式的，但很少。

目录转换/网站转换/批量转换

1）目录转换

在“文件、网页、目录、网站简繁体转码”中转换对象为“一个目录/网站”即可。设置此选项时，默认将“包含子目录”，此时该目录下的所有文本文件都会被转换。您不必担心目录下的非文本文件。软件会自动识别文本文件，非文本文件会被原样复制到目标目录下的相应位置。

2）网站转换

一个网站一般就是一个目录。一般而言，如果网站中有多种语言版本，一般一个语言版本会转放为一个目录，如果你有简体版，需要繁体版，只需将此简体目录转换为繁体即可。如有繁体想转简体，反之操作即可。

传统的简繁体网页的编码一般分别是 GBK（即 GB2312）和 Big5。现在的潮流则时统一采用 UTF-8 编码。您转换时，选对源编码，再根据需要设置目标编码为 GBK（简体）、Big5（繁体）或 UTF-8（简体或繁体），并进行转换即可。

网站一般都有多层目录，除了文本文件、网页文件，还有图片等其它格式文件，只要采用默认设置，软件会自动把所有各级目录下内容一次性进行转换，并会将除文本文件、网页文件外的图片、执行文件、压缩文件等各种其它格式文件自动复制到目标目录中的相应位置。

3）批量转换

如果您有多个文件或目录需要成批进行转换，可以将其复制到一个指定目录下，再对此目录统一进行转换即可。

【特殊格式文件的简繁转换】

特殊的文件格式，如 Word、Excel、PowerPoint、Access、WPS、FrameMaker、Trados 等它们的标准格式并不是上面写的这些编码，但幸运的是，这些著名产品的开发商都意识到标准编码格式的重要性，他们的较新版本都提供了可以导出为 XML（即 UTF-8）或 GBK、Big5 等编码文件的功能，行业的主流趋势是绝大多数重要软件都会支持导出为 XML（UTF-8）格式。

转换特殊格式文件时，一般的流程是先将其导出为 XML（即 UTF-8）格式，进行文件或目录的 UTF-8 简繁转换（即“**简体中文（UTF-8）**”<->“**繁体中文（UTF-8）**”），需要的话，也可以以 UTF-8 或 Unicode 格式的编辑器打开，并进行剪贴板 UTF-8 或 Unicode 简繁转换。

PDF

PDF 文档比较特殊，它并不是一种可编辑的文档格式。PDF 一般是由 Acrobat Professional 或具有类似功能的软件从纯文本文件、图片文件或上述主流特殊格式文件（如 Word、Excel、FrameMaker 等）转换而成。一般你不能直接对 PDF 文档进行简繁转换，但你可以对用于制作该 PDF 文档的源文档先进行简繁转换，然后将转换形成的文档再制成 PDF 即可。

Word

需要 Word 2003 以上版本。

步骤：

1） 另存为“XML 文档”

2） 对该文档进行文件转换（“简体中文（UTF-8）”<->“繁体中文（UTF-8）”）

3） 用 Word 打开后（双击即可）另存为 Word 文档就行。

Excel

需要 Excel 2003 以上版本。

步骤：

1） 另存为“XML 表格”

2） 对该文档进行文件转换：“简体中文（UTF-8）”<->“繁体中文（UTF-8）”

3） 用 Excel 打开后另存为 xls 文档就行。

PowerPoint

需要 PowerPoint 2007 以上版本。

步骤：

1） “另存为”->“其它格式” -> PowerPoint XML 演示文档（*.xml）

2） 对该文档进行文件转换：“简体中文（UTF-8）”<->“繁体中文（UTF-8）”

3） PowerPoint 打开后另存为 ppt 文档就行。

Access

需要 Excel 2003 以上版本。

Access 的简繁转换需要对每个对象分别进行简繁转换。

步骤：

1） 选中一个对象（如一个“表”、“查询”、“窗体”等），并“导出”，导出为“XML (*.xml)”
2） 选择“数据（XML）”+“数据架构（XSD）”，并完成导出
4） 对导出的这两个文件进行简繁转换：“简体中文（UTF-8）”<->“繁体中文（UTF-8）”
3） 注意，如果两项同时要保存，且文件名为中文名或包含中文，您需要同时对文件名进行简繁转换，这是因为文件中要调用文件名，而文件中的文件名已经进行过简繁转换了。如果您使用简体中文操作系统，需要对文件名进行简繁转换，只需要在“选项”中设置“转换。在 GBK 简繁之间转换（仅限于简体中文系统）”。

WPS

步骤：

1） 另存为“中文办公软件文档格式（*.uof;*.xml）”
2） 对该文档进行文件转换（“简体中文（UTF-8）”<->“繁体中文（UTF-8）”）
3） 用 WPS 打开后（双击即可）另存为 WPS 或其它需要的格式即可。

FrameMaker

如您用的是简体中文操作系统，另存为“MIF (*.mif)”，这样存下来的是 GBK 格式文件。然后你对此文件进行 “简体中文（GBK）”->“繁体中文（GBK）”文件转换即可，转换出的文件虽然是 GBK 格式，但在繁体中文和英文系统下都能正常阅读。

FrameMaker 里也可以保存为 XML 文件，也可按（1）先另存为 XML（2）进行“简体中文（UTF-8）”<->“繁体中文（UTF-8）”转换的方式处理，但似乎另存为 XML 时会丢失东西，所以不推荐此流程。

Trados

1）TM 的转换

方式 1：采用对“Translators' Workbench (*.txt)”导出文件进行转换。

以英译简转英译繁体为例（以 Trados 7.0 为例）

1.1：先将将 TM 导出为 txt：File->Export->OK->保存类型：Translators' Workbench 7.x (*.txt)->如取名为“mytmchs.txt”

1.2：转换，用《简繁大师》进行简繁转换：“简体中文（UTF-8）”->“繁体中文（UTF-8）”->“mytmcht.txt”

1.3：新建空白英译台湾繁体 TM，如 mytmcht_temp，在 Word 里随便翻译两个词，如 software->軟體，hardware->硬體，然后导出，观察一下导出文件和和简体的导出文件有什么区别。差别如下：

1）文件的开头部分有差别：字体上有区别，还有些其它差别

2）在翻译单元中，由简体转过来的繁体版本，目标单元的标记是“<Seg L=ZH-CN>”，而直接生成的繁体版中的标记是“<Seg L=ZH-TW>”。源文的标记正好相同，都是“<Seg L=EN-GB>”（表示源语言都是英国英语）。

1.4: 用查找替换的方法把刚转换出的繁体文件的标记改成繁体应有的标记：

目标语：“<Seg L=ZH-CN>”->“<Seg L=ZH-TW>”

源语正好相同，不必改，如果不同（比如一个是美国英语，一个是英国英语），也可以一起改了

1.5：用改好的繁体条目覆盖过渡 TM 的条目部分。保存

1.6 再次按 1.3 格式新建空白英译台湾繁体 TM，这次玩真的了，生成后，用 1.5 步最后形成的 txt 文件导回（File->Import->OK->选择刚才保存的 txt）并保存。

完成了，可以测试一下看看是否合格。
成功了。

方式 2：采用对“TMX 1.4b (*.tmx)”导出文件进行转换。
说明：TMX 是业内的 TM 标准格式。所以对此文件进行转换具有普遍意义。参考方式 1）处理如下
以英译简转英译繁体为例（以 Trados 7.0 为例）
1.1：先将将 TM 导出为 TMX 1.4b (*.tmx)：File->Export->OK->保存类型：“TMX 1.4b (*.tmx)”->如取名为“mytmchs.tmx”
1.2：转换，用《简繁大师》进行简繁转换：“简体中文（Unicode）”->“繁体中文（Unicode）”->“mytmcht.tmx”
1.3：新建空白英译台湾繁体过度 TM，如 mytmcht_temp，在 Word 里随便翻译两个词，如 software->軟體，hardware->硬體，然后导出，观察一下导出文件和和简体的导出文件有什么区别。差别如下：
1）文件的开头部分有差别：字体上有区别，还有些其它差别
2)在翻译单元中，由简体转过来的繁体版本，目标单元的标记是“<tuv xml:lang="ZH-CN">”，而直接生成的繁体版中的标记是“<tuv xml:lang="ZH-TW">”。源文的标记也不同（因为在这个例子中我们改用了美国英语，一个是“<tuv xml:lang="EN-GB">”，另一个是“<tuv xml:lang="EN-US">”。
1.4: 用查找替换的方法把刚转换出的繁体文件的标记改成繁体应有的标记：
源语：“<tuv xml:lang="EN-GB">”->“<tuv xml:lang="EN-US">”
目标语：“<tuv xml:lang="ZH-CN">”->“<tuv xml:lang="ZH-TW">”
1.5：用改好的繁体条目（<body>和</body>之间的部分）覆盖过渡 TMX 的相应部分。
1.6 再次按 1.3 格式新建空白英译台湾繁体 TM，这次玩真的了，生成后，用 1.5 步最后形成的 txt 文件导回（File->Import->OK->选择刚才保存的 tmx）并保存。
完成了，可以测试一下看看是否合格。
成功了。

2）双语 ttx 的转换
一般情况下不需要直接转换 ttx 文件，推荐的做法是转换 TM，然后用 TM 来转换相关文档。
当然，多一种转换技能也不是什么坏事。由于相似的词条在 TM 中只存为一条，所以就算您有了 TM 也并不能保证对相关稿稿实际 100% 完美转换，所以对于对品质要求较高的客户，转换精心处理过的 ttx 也有其不可替代的优势。
对比一下简体中语言 ttx 和刚才转换的繁体中文 ttx。它们都是 Unicode 文件。
区别如下：
SourceLanguage: "EN-GB" -> "EN-US" (全部替换，这只是我们的问题，如果原来就是美国英语，则不必改)
TargetLanguage："ZH-CN" -> TargetLanguage="ZH-TW" (全部替换)
TargetDefaultFont："幼圓" -> "MingLiU" (全部替换)
保存后用刚才做出来的繁体 TM 打开看看。看起来不错。
clean 一下（clean 时选择如果不想更新 TM，可以选择“Don't update”，否则选择“Update TM”），Workbench 会问你“如果您计划将此演示文稿发送回给原作者，可以将您的更改标记为修订。当作者将其与原始演示文稿合并时，修订标记将显示，指示您编辑过的内容。保

存此文件时是否自动添加这些附加信息？”一般情况下无此需要，选择“否”。
成功了。

3）双语 Word 文档的转换

这里的双语 Word 文档指用 Trados 转换后未 clean 过的 uncleaned 版文档。
与双语 ttx 一样，您可以直接转换 TM，再用转换出来的 TM 处理 Word 原文，得到新的繁体版双语 Word 文件。但这样有时会不能保证对所有词条 100% 准确翻译。如果您需要 100% 准确翻译，可以考虑直接转换双语 Word 文件。
3.1 在英译简体中文项目中，有了简体中文版双语 Word 文件，将该双语文件另存为“XML 文档 (*.xml)”，该文件的格式是 UTF-8，对该文件进行文件简繁转换：“简体中文（UTF-8）” -> “繁体中文（UTF-8）”。
3.2 做一个英译台湾繁体中文，试做一个简体的繁体中文版双语 Word 文件，将该双语文件另存为“XML 文档 (*.xml)”，该文件的格式是 UTF-8，打开与上面的转换结果进行对比。
对比结果如下：
1）用"幼圓"的地方，改成用"MingLiU"（全部转换），用"幼圓"的地方，改成用"MingLiU"（全部转换），
//2）用"宋體"的地方，只要用"宋体"即可（照理需要转换成简体，不过 Word 也识别繁体字，所以先不管了）
2）用"宋體"的地方，只要用"宋体"即可。台湾繁体中也有宋体，但在简体中文 Word 中，台湾的这个“宋体”在简体中文 Word 中的名称就叫“宋体”。照理简体中文版 Word 也识别繁体字，不过我们的实验显示字体中的"宋體"如果不转换成"宋体"，以后"宋體"仍会保留。不知后果如何，多一事不如少一事，就算了。
3）有些说明用了不该用的繁体，暂时不管了。Word 会自动调整的。
4）用 Word 打开，看起来没什么明显问题，另存为 doc 文件。
5）再次另存为 XML 文档，看看上面没改的繁体现在如何了。
看样子结果可以接受。

【出现乱码的原因及解决方案】

出现乱码的两大类原因

1）或未采用适当的方式打开或显示相应编码的文本。

2）进行简繁转换时选错了转换源编码或转换方向，主要是选错了源编码。

现在分别看看两种情况。

1）未采用适当的方式打开或显示相应编码的文本，例如：

GBK

在简体中文操作系统中用一般的编辑器都可以打开。在其它语言操作系统中可以用《简繁大师》打开 GBK 文件或查看 GBK 文本（显示语言设置为“简体中文（GBK）”）。

Big5

在繁体中文操作系统中用一般的编辑器都可以打开。在其它语言操作系统中可以用《简繁大师》打开 Big5 文件或查看 Big5 文本（显示语言设置为“繁体中文（Big5）”）。

UTF-8/Unicode/Unicode BE

在一般操作系统中支持相应编码的编辑器可以打开查看，有时在打开时需要专门指定相关编码（如打开 UTF-8 文件时，有时需要指定编码为 UTF-8）。

在普通 Windows 操作系统中，从网页或 Word 上复制出来的文本，其编码为 Unicode。《简繁大师》可以打开 UTF-8，但显示出的是“乱码”，如果您想了解其内容，可以在编辑区内进行转换：“简体中文（UTF-8）” -> “简体中文（GBK）”（此转换也适用于内容是繁体的情况）。

《简繁大师》不能直接打开或显示 Unicode 或 Unicode BE，在简体中文或繁体中文系统中，源码为 Unicode 的文本直接贴入简繁大师时，会自动分别转为 GBK 和 Big5。在英语或其它语言系统中，则不会转换。当然您也可以自己先在剪贴板中转换为 GBK 或 Big5 后再贴入，就可以了解其内容。

如果您将剪贴板中的内容贴入《简繁大师》时显示的是很多英文问号：

??????????????

则表示系统未能帮您自动转换，碰到这种情况，如果想在《简繁大师》中正确显示，请自行进行剪贴板“简体中文（Unicode）” -> “简体中文（GBK）”或“繁体中文（Unicode）” -> “繁体中文（Big5）”后再贴入。工具条上的按钮  和  可用于进行此类转换。

 = 剪贴板转换：“简体中文（Unicode）” -> “简体中文（GBK）”

 = 剪贴板转换：“繁体中文（Unicode）” -> “繁体中文（Big5）”

另外两个按钮  和  的功能正好与上面相反。

 = 剪贴板转换：“简体中文（GBK）” -> “简体中文（Unicode）”

 = 剪贴板转换：“繁体中文（Big5）” -> “繁体中文（Unicode）”

另一种特殊情况是未安装适当字体。
一般碰到这种情况操作系统或简繁大师会提示您安装字体。如果您能在浏览器中访问简体中文和繁体中文的网页，表示字体已安装。

2）你进行转换时选错了转换方向，主要是选错了源编码，比如源文件是 UTF-8，结果你把它当成 GBK 或 Unicode 进行转换，那转出来的自然是乱码。

【命令行功能】

简繁大师标准版提供了命令行功能，如果您非常熟悉命令行操作，或需要在您自己的程序中调用简繁大师的命令行功能，可以使用此功能。具体细节请参帮助文件附录中的相关说明。