CUC_ParaConc 使用说明

1 概述

CUC_ParaConc（中国传媒大学平行语料检索）软件是一个免费的绿色软件，设计的目的是减轻研究者的劳动量。软件主要用于检索双语、多语平行语料，支持对 Unicode、UTF8、ANSI 等编码的纯文本语料检索，支持多个国家的平行语料检索，例如：汉语、英语、法语、俄语、韩语、日语、泰语等。多语检索可以实现 1 对 16 的平行语料，即一个原文对齐的 1～16 个国家的译文。

软件从 0.3 版本开始，增加了较多功能：

（1） 英汉双语界面，并可以自己修改界面，把界面翻译成任意一种语言；

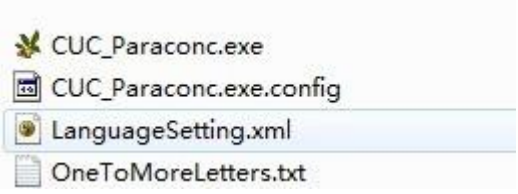
（2） 排序功能；

（3） 对于双语保存在一个文本中的平行语料，可以自动识别其对齐形式；

（4） 关键词居中变色功能

（5） 多语检索，由 1 对 8 检索，增加到 1 对 16 检索。

2 软件主要组成部分



软件共由四个部分组成：

（1） CUC_Paraconc.exe ：可执行文件

（2） CUC_Paraconc.exe.config： 配置文件

Config 文件是配置文件，用 txt 打开后，可以修改里面的配置，比如：

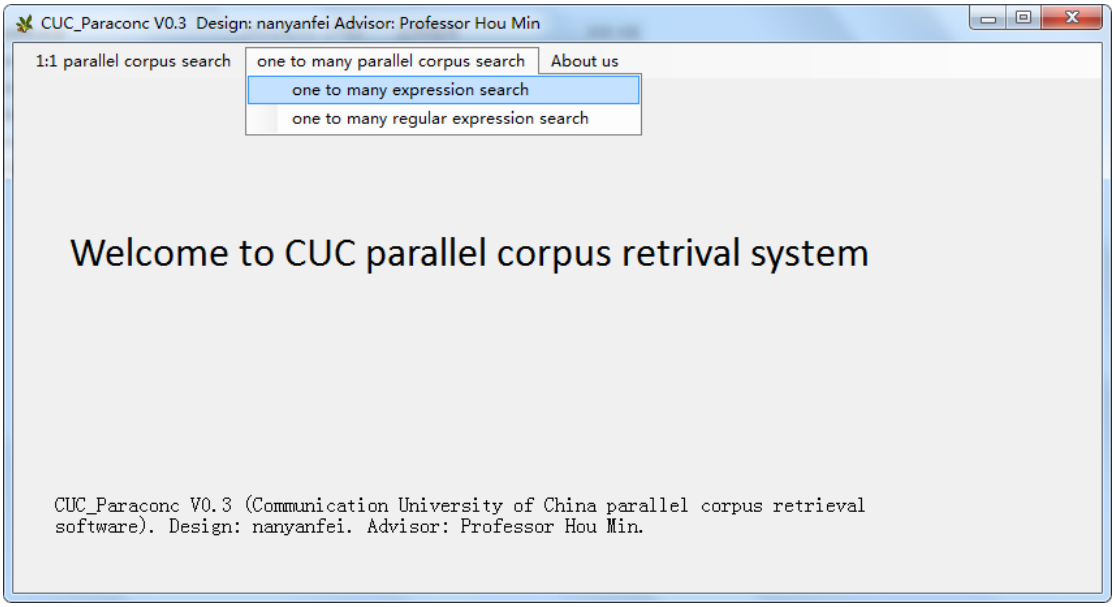
```
<setting name="languageSetting" serializeAs="String">
    <value>English</value>
</setting>
```

把 “<value>English</value>”中间的 Englisth，修改成“Chinese”，界面就会以

中文的形式呈现。其他的参数可以根据需要进行修改，也可以保持默认。Config 文件内对每个参数有简单的注释。

（3）LanguageSetting.xml：界面语言设置

XML 里面是界面及提示语言的汉英双语平行语料，比如：




```
<para id="4" name="menuRegEx">
    <sent id="Chinese" cont="双语正则式检索"></sent>
    <sent id="English" cont="Bilingual regular search"></sent>
  </para>
```

可以修改 cont 里边的汉英界面译文：

双语正则式检索

Bilingual regular search

修改后，再次打开软件，界面就可以改变。您可以根据需要进行翻译，可以翻译成任意一种语言。

（4）OneToManyLetters.txt

OneToManyLetters.txt 里面是 1 对 16 多语检索的时候，17 种语言的字母表，默认都为英语。您可以根据需要，修改成正在检索语言的所有字母，包括大小写。可以从下面的菜单中进入该检索窗口（模块）：

2　一对一双语检索使用说明

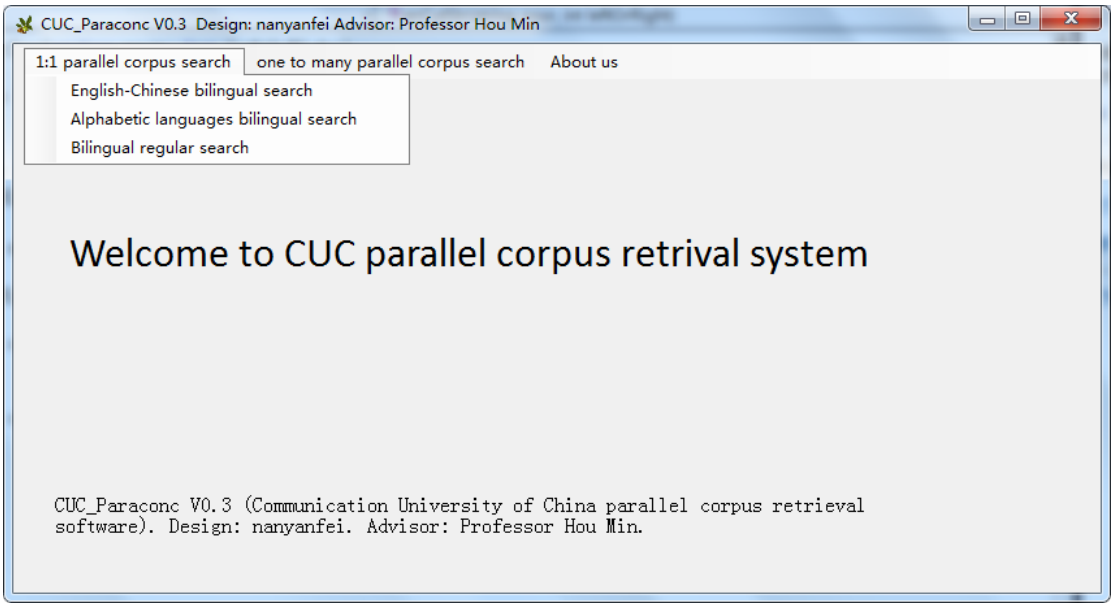
2.1　一对一双语检索包括三个检索窗口：

（1）英汉双语检索：用来检索汉语和英语平行语料

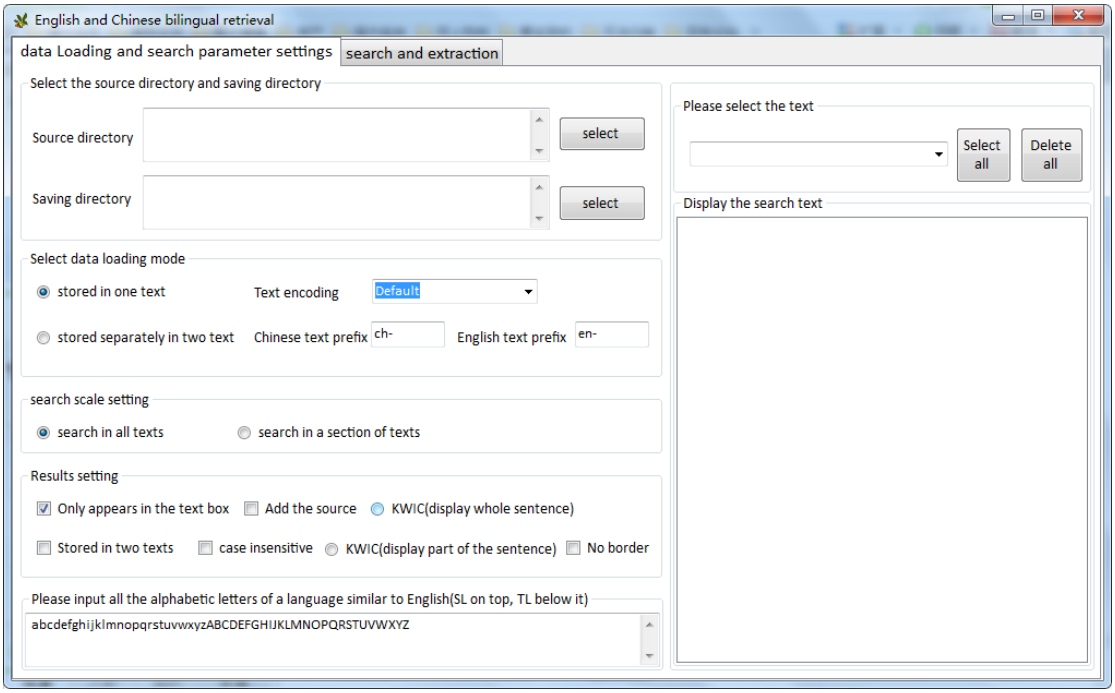
（2）拼音文字检索：用来检索拼音文字双语平行语料，比如英语和法语。同时也可以检索分词后的汉语，您只需要中字母中间加入一个空格。

（3）正则式检索。

进入菜单如下图所示：



2.2 英汉双语检索窗口使用说明



2.2.1 加载语料

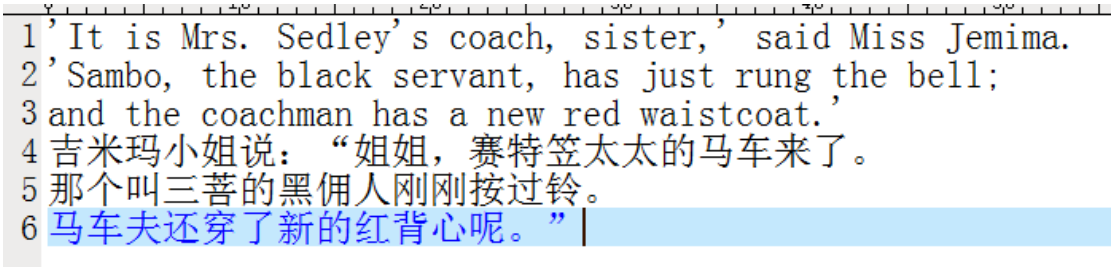
2.2.1.1 双语保存在一个文本中

首先要选择源文本目录和保存文本目录。从 0.3 版本开始，对于双语保存在一个文本中的平行语料，软件能自动识别四种对齐形式，汉语在前，汉语在后，汉语整体在前，汉语整体在后，并可以把四种形式的文本乱在一起检索。注意一点，每个文本的形式必须是一致的，不是说一个文本中存在四种形式，而是每个文本一种形式，比如：下面的文本形式

A 文本是英语在前，汉语在后：

```
1 Report on the Work of the Government (2000)
2 2000年政府工作报告
3 Delivered at the Third Session of the Ninth Nationa
4 ——2000年3月5日在第九届全国人民代表大会第三次会议_
5 Zhu Rongji Premier of the State Council
6 国务院总理朱镕基
```

B 文本是英语整体在前，汉语整体在后



A、B 两个文本可以乱在一起进行检索。

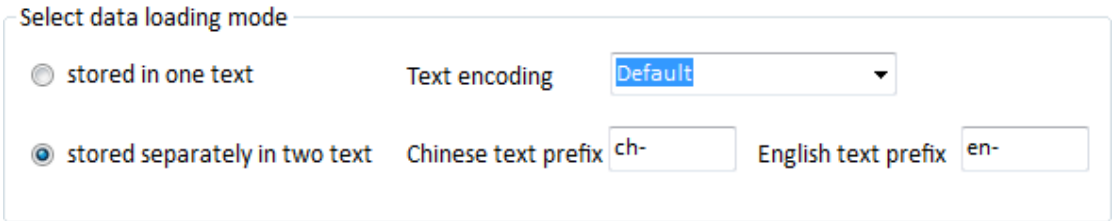
2.2.1.2 双语保存在两个文本中

一对一双语对齐的语料，另一种保存形式是双语分开保存在两个文本中，这种情况文件的命名一定要有规律，否则无法检索。比如有"中文、英文"两个文本语料，那么可以这样命名它们"ch-test.txt，en-test.txt"，"ch-"就是中文文件名的前缀，"en-"就是英文文件名的前缀；或者命名为"test.txt，en-test.txt"，这里中文文件名的前缀为空，但是英文文件名为"en-"；或者"ch-test.txt，test.txt"， 这里英文文件名的前缀为空，但是中文文件名为"ch-"，也就是说，其中中文或者英文两个文本中的一个一定要加一个前缀，或者两个都加前缀，"test"部分是必需相同的。也就是说，文件由两部分组成：前缀+文件名。比如现在有汉语《三国演义》的第一章、第二章、第三章、第四章，同时有英文译本的《三国演义》的第一章、第二章、第三章、第四章，下面是命名方法举例：

章节 / 对齐语言	第一章	第二章	第三章	第四章
	前缀+相同部分	前缀+相同部分	前缀+相同部分	前缀+相同部分
汉 语	ch-sgyy1.txt	ch-sgyy2.txt	ch-sgyy3.txt	ch-sgyy4.txt
英 语	en-sgyy1.txt	en-sgyy2.txt	en-sgyy3.txt	en-sgyy4.txt

前缀可以自己定义，不一定是软件默认的"ch-""en-"，也可以用中文前缀。比如："原著""译著"，或者"中""英"等。

双语分开保存的语料加载的时候，要保证复选框"双语分开保存在两个文本中"为打勾状态。如下图所示：



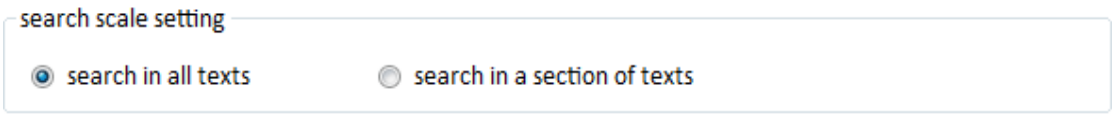
软件提供了输入中文与英文前缀的文本框，两者必需有一个不为空，或者两个都不为空（建议使用）。选择的源文本目录下可以包括多个文件夹，但是对齐的两个语料文本一定要在同一个文件夹中。

2.2.2 参数设置

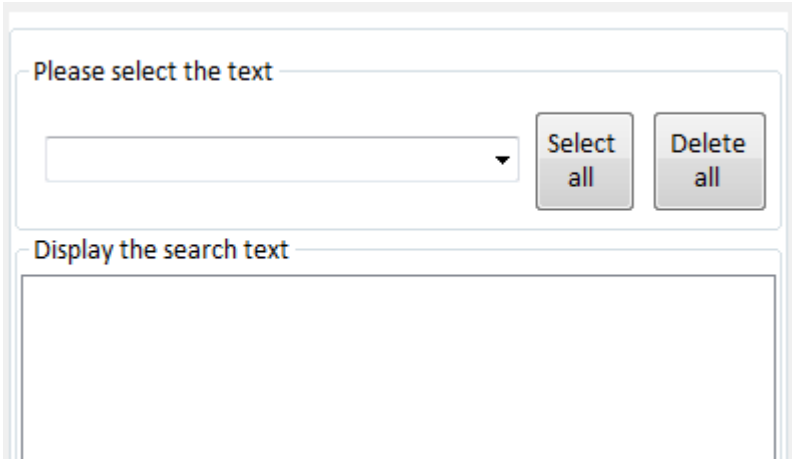
2.2.2.1 编码设置

在语料的加载中还有一个下拉框：语料编码方式，如上图所示，初始状态是“系统默认”编码，您可以根据语料的存储形式进行选择，如果选择错，可能会出现乱码，或者检索结果为空的。

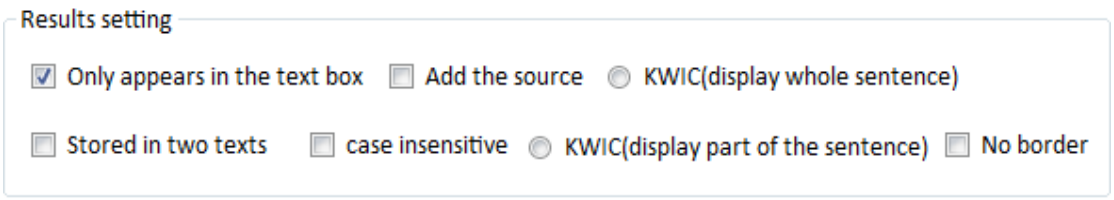
2.2.2.2 检索规模设置



可以设置成在所有语料中检索和在部分语料中检索，在部分语料中检索，可以通过下图所示的界面进行选择，如果选择错误，可以通过双击删除该文本。

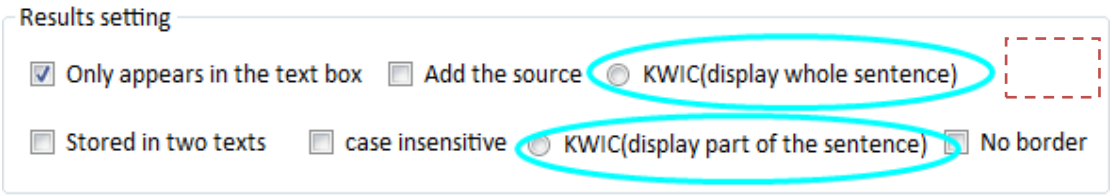


2.2.2.3 检索结果设置

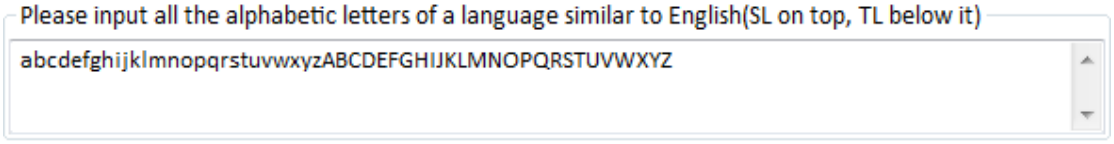


如果没有特殊要求，可以保持默认，这样检索结果会显示在软件的结果显示框中。如果要直接保存到文本中，可以去掉“检索结果只显示在文本框中”复选框的勾号。去除勾号后，“检索结果分开保存在两个文本中”这个复选框恢复为可选状态，如果不选择该复选框，语料会保存为一个文本中，选中后，语料会分开保存为两个文本：中文结果、英文结果。保存的文本在您选择的保存到目录下。

下图所示关键词居中显示，只能二选择一，如果想取消选择，可以双击右边虚线框表示的空白处。No border 只适用于部分显示的时候。

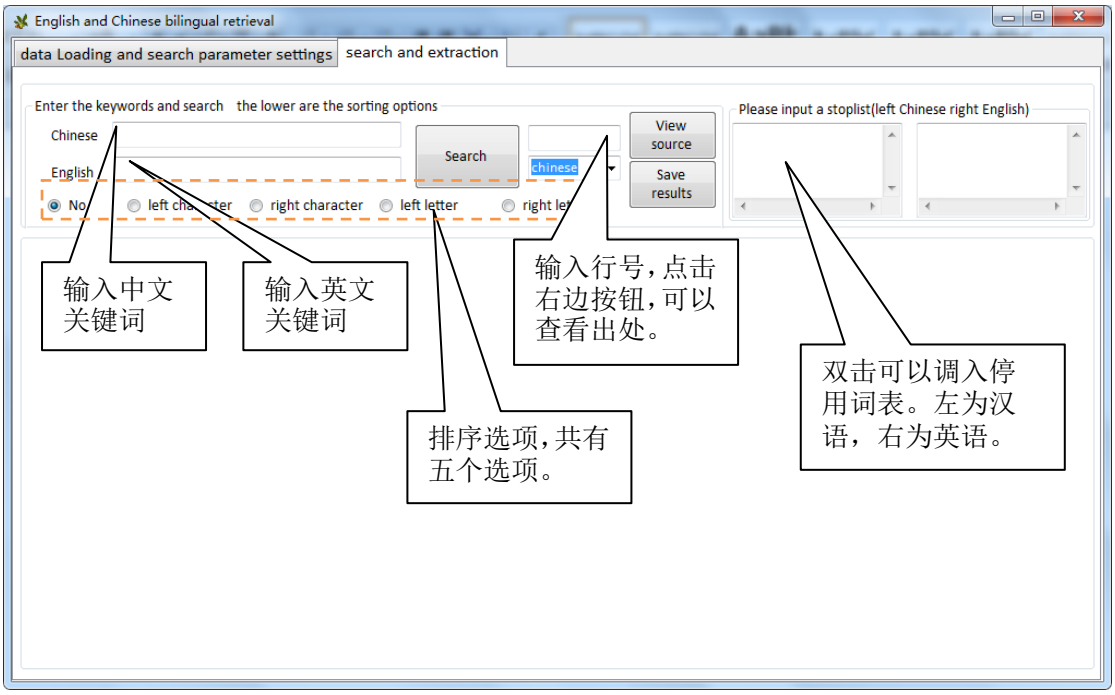


2.2.2.4 字母设置

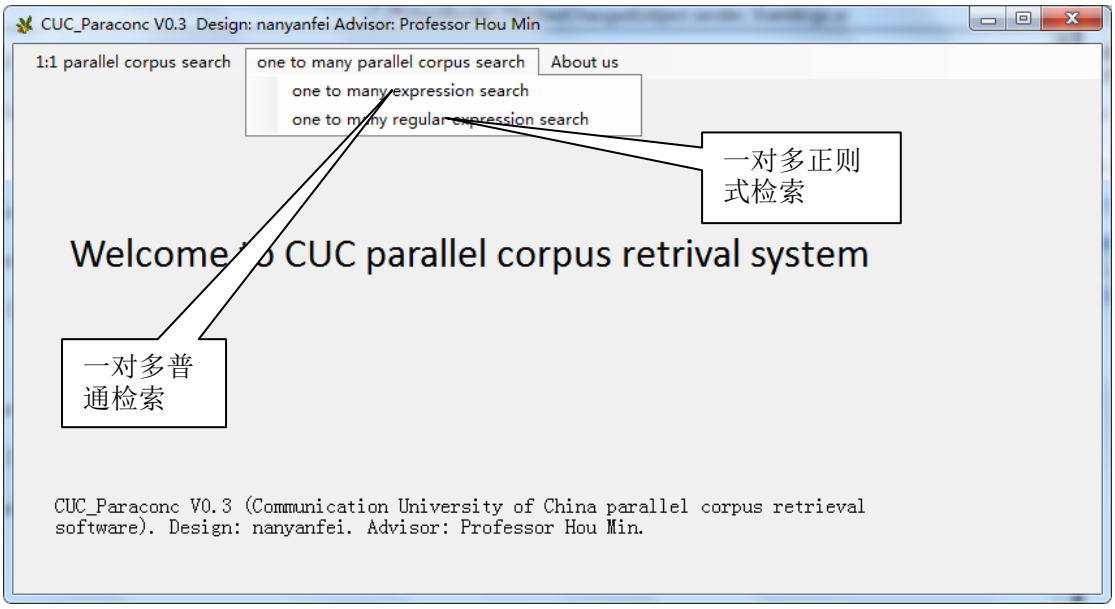


可以在上图所示的文本框中输入要检索的某种语言（拼音文字）的所有字母，包括大小写。

2.2.3 一对一检索


English and Chinese bilingual retrieval
data Loading and search parameter settings
search and extraction
Enter the keywords and search the lower are the sorting options
Chinese
English
Search
chinese
View source
Save results
Please input a stoplist(left Chinese right English)
right character
left letter
输入中文关键词
输入英文关键词
输入行号，点击右边按钮，可以查看出处。
排序选项，共有五个选项。
双击可以调入停用词表。左为汉语，右为英语。

3 一对多双语检索使用说明

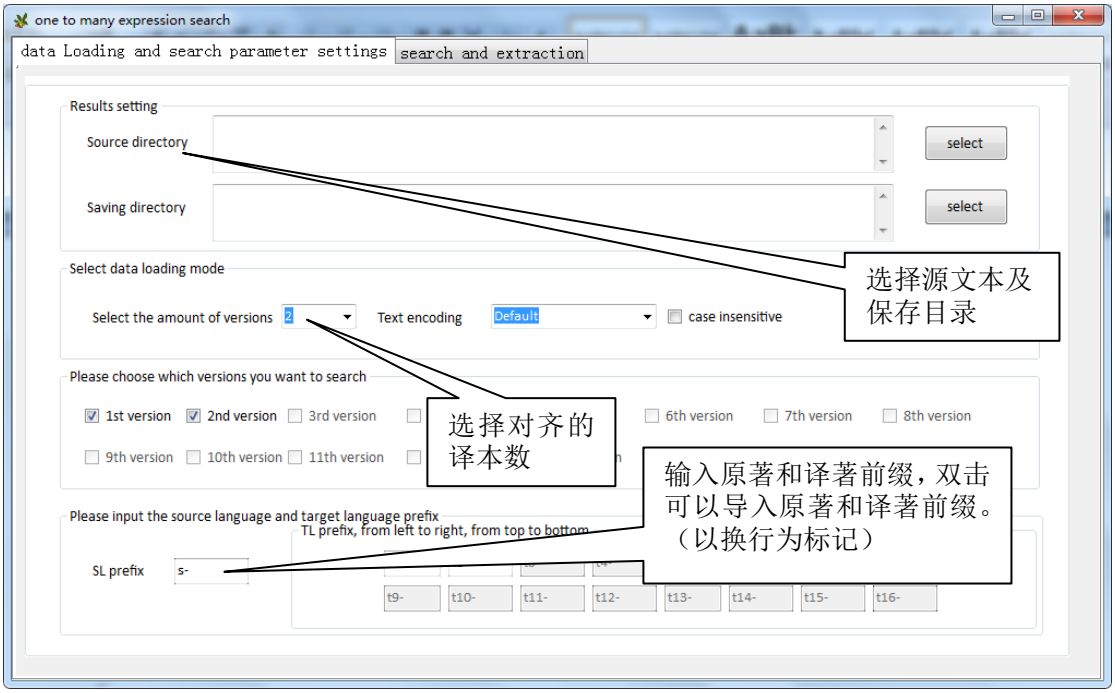


该菜单下有两个子菜单，

一对多普通检索，与前面提到的 OneToMoreLetters.txt 文件相关，需要在该文件中设置需要检索的语言所有字母，包括大小写。

3.1 载入语料

一对多平行语料检索窗口可以实现任意一种语言语料与另外 1～16 对齐的任意语言语料的检索。我们用中、英、法、德、日、韩、阿拉伯、俄等多种语言进行了测试，都能正常检索和显示。该模块不再区分哪一种语言，只分“原著”和“译著”，原著是“一”，译著是“多”，可以选择 1～16 种相同或者不同的语言语料。一对多有两种情况，一是双语语料语料，即一个原著的多个相同语言的译本，比如《三国演义》的四个英文译本，另一种是多语对齐，比如《三国演义》的英、日、韩、泰等多个译本。因此在检索之前，先要区分的就是原著、译著，原著是“一”，译著是“多”，一对多检索的概念来源于此。



首先选择对齐的译本数：默认是 2，最大极限是 16，接着选择语料的编码方式，也就是您自己的 txt 文本的编码，多国语言一般是 Unicode 编码或者是 UTF8 编码。然后输入原著和译著的前缀，注意，原著和译著都要输入前缀。前缀可以是任意字符，数字，英文字母，汉字等等。下面我们照样以前面提到的《三国演义》四个章节的译文为例子进行说明，现在假设有《三国演义》的 8 个外文译本，它们分别是：拉丁译文、英译文、德译文、荷译文、俄译文、爱沙尼亚译文、波译文，我们现在自定义原文与译文的前缀如下：

原文《三国演义》前缀为	0
拉丁译文前缀为	1
英译文前缀为	2
德译文前缀为	3
荷译文前缀为	4
俄译文前缀为	5
爱沙尼亚译文前缀为	6
日译文前缀为	7
波译文前缀为	8

前缀定义为数字，需要记住 1～8 的数字分别代表哪一部译著，但如果您的语料多的话，批量命名文件的软件就可以发挥作用。如果您觉得不习惯，可以自

定义为其它的前缀。我们先以这种前缀为例子来给《三国演义》的四个章节的译本命名。

章节 / 对齐语言	第一章 前缀+相同部分	第二章 前缀+相同部分	第三章 前缀+相同部分	第四章 前缀+相同部分
汉语(原文)	0sgyy1.txt	0sgyy2.txt	0sgyy3.txt	0sgyy4.txt
拉丁译文	1sgyy1.txt	1sgyy2.txt	1sgyy3.txt	1sgyy4.txt
英译文	2sgyy1.txt	2sgyy2.txt	2sgyy3.txt	2sgyy4.txt
德译文	3sgyy1.txt	3sgyy2.txt	3sgyy3.txt	3sgyy4.txt
荷译文	4sgyy1.txt	4sgyy2.txt	4sgyy3.txt	4sgyy4.txt
俄译文	5sgyy1.txt	5sgyy2.txt	5sgyy3.txt	5sgyy4.txt
爱沙尼亚译文	6sgyy1.txt	6sgyy2.txt	6sgyy3.txt	6sgyy4.txt
日译文	7sgyy1.txt	7sgyy2.txt	7sgyy3.txt	7sgyy4.txt
波译文	8sgyy1.txt	8sgyy2.txt	8sgyy3.txt	8sgyy4.txt

以上只是其中的一种定义前缀的方法，您还可以用汉字定义，比如我们对原文和译文定义如下的前缀：

原文《三国演义》前缀为	原文
拉丁译文前缀为	拉丁
英译文前缀为	英
德译文前缀为	德
荷译文前缀为	荷
俄译文前缀为	俄
爱沙尼亚译文前缀为	爱沙
日译文前缀为	日
波译文前缀为	波

针对如上的前缀，那么我们对原文和译文的文本的命名要作如下修改：

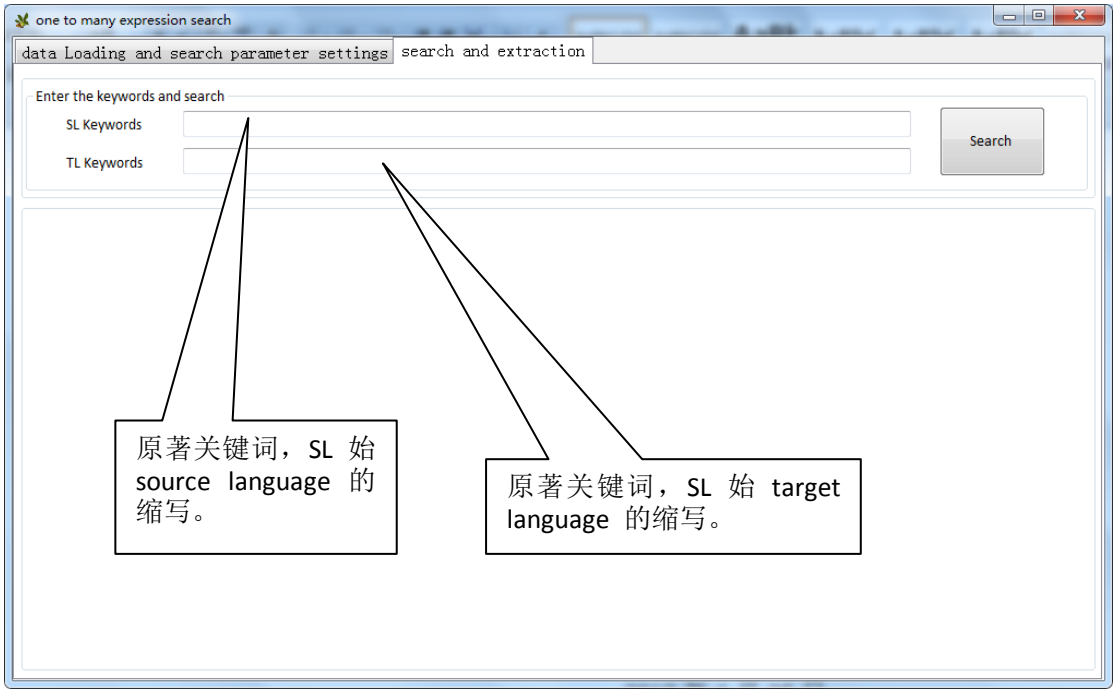
章节 / 对齐语言	第一章 前缀+相同部分	第二章 前缀+相同部分	第三章 前缀+相同部分	第四章 前缀+相同部分
汉语(原文)	原文 sgyy1.txt	原文 sgyy2.txt	原文 sgyy3.txt	原文 sgyy4.txt
拉丁译文	拉丁 sgyy1.txt	拉丁 sgyy2.txt	拉丁 sgyy3.txt	拉丁 sgyy4.txt
英译文	英 sgyy1.txt	英 sgyy2.txt	英 sgyy3.txt	英 sgyy4.txt
德译文	德 sgyy1.txt	德 sgyy2.txt	德 sgyy3.txt	德 sgyy4.txt
荷译文	荷 sgyy1.txt	荷 sgyy2.txt	荷 sgyy3.txt	荷 sgyy4.txt
俄译文	俄 sgyy1.txt	俄 sgyy2.txt	俄 sgyy3.txt	俄 sgyy4.txt
爱沙尼亚译文	爱沙 sgyy1.txt	爱沙 sgyy2.txt	爱沙 sgyy3.txt	爱沙 sgyy4.txt
日译文	日 sgyy1.txt	日 sgyy2.txt	日 sgyy3.txt	日 sgyy4.txt
波译文	波 sgyy1.txt	波 sgyy2.txt	波 sgyy3.txt	波 sgyy4.txt

当然您还可以用英文前缀来命名，在此就不再举例了。

如果知道了文件命名的规则后，加载语料就容易多了。还有一个选项要注意的就是可以选择在多个译本中的哪一部，或者是哪几部译本中进行查找。希望这个功能可以更加方便大家。

特别提醒：多语检索文本编码一般是 Unicode 别忘记了选择编码方式。

3.2 检索



程南昌 2013-6-23

于中国传媒大学