SDAU-ParaConc 使用说明

山东农业大学外国语学院 葛晓帅

版本 0.0.3.01

软件介绍：

该软件为免费软件。

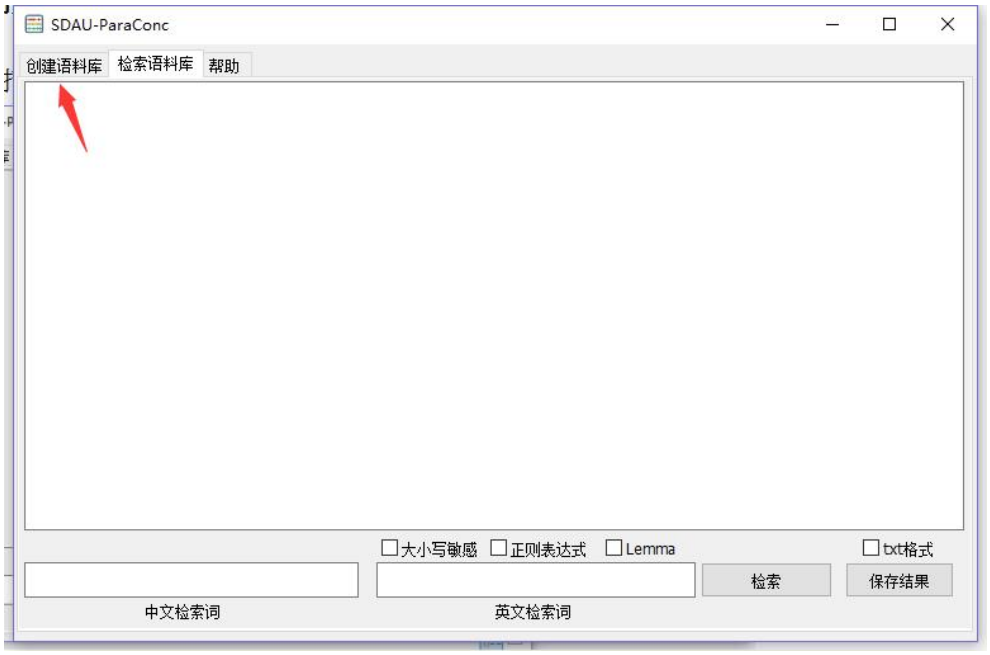
该软件为平行语料库检索软件，支持文本文件（TXT）和 tmx 文件。

支持任意编码。中文文本不需要分词。

软件功能：

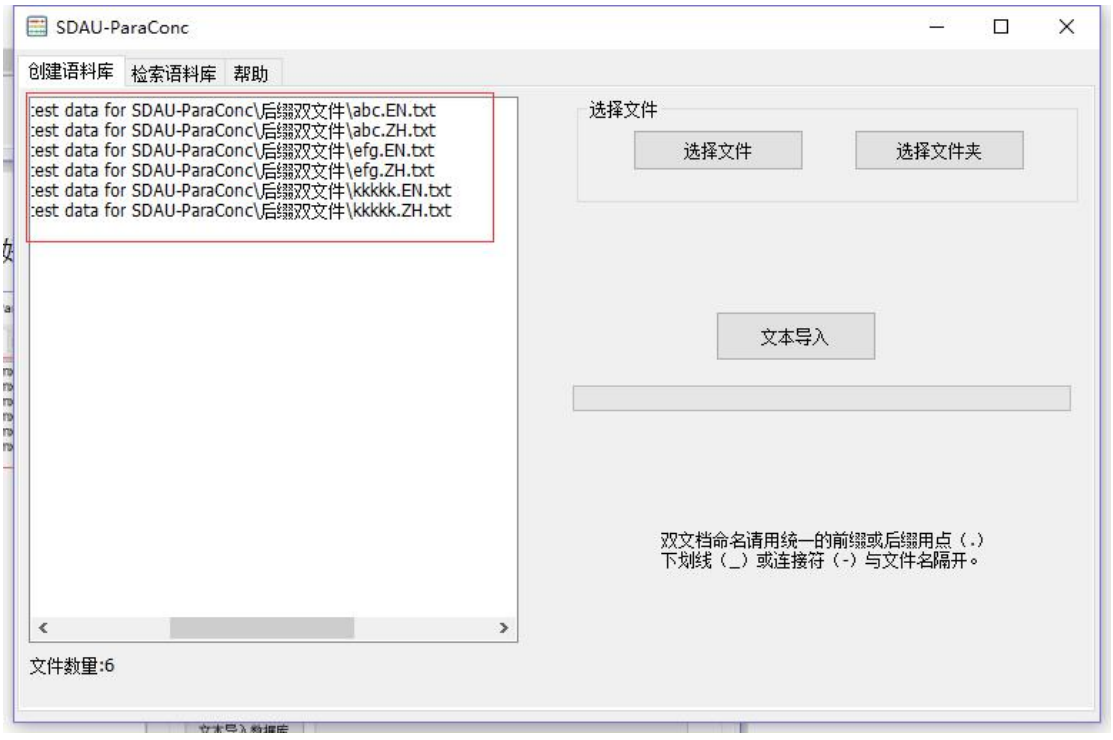
一、创建语料库

1 双击打开软件后，点击**创建语料库**选项卡



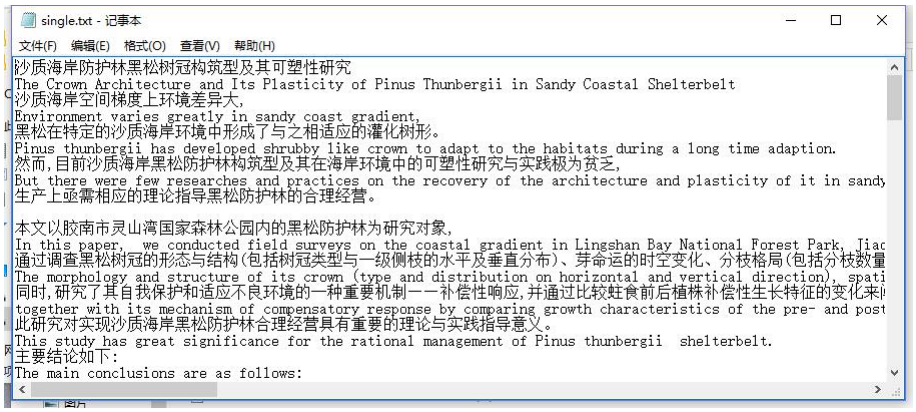
2 选择您要检索的文本，您可以通过点击“选择文件”或“选择文件夹”进行选择：

3 选择好的文件会出现在左侧列表框中：

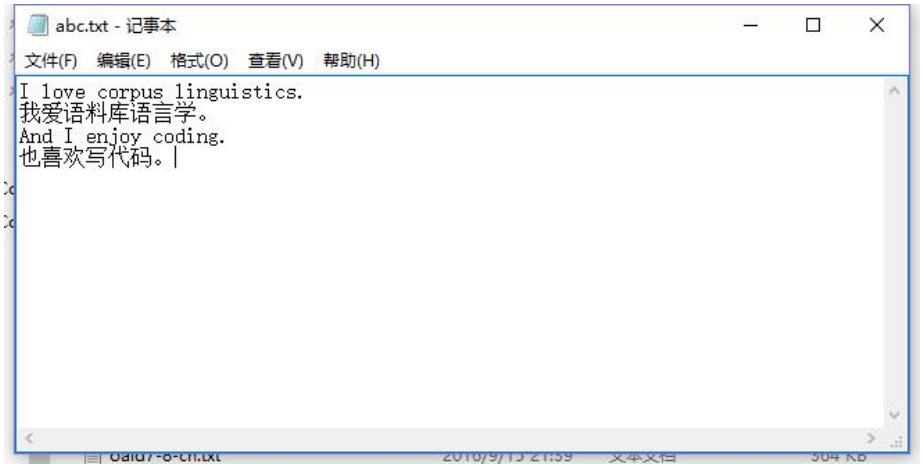


4 关于导入文本文件类型的说明：
软件支持单文件对齐语料，双文件对齐语料以及 tmx 文件。

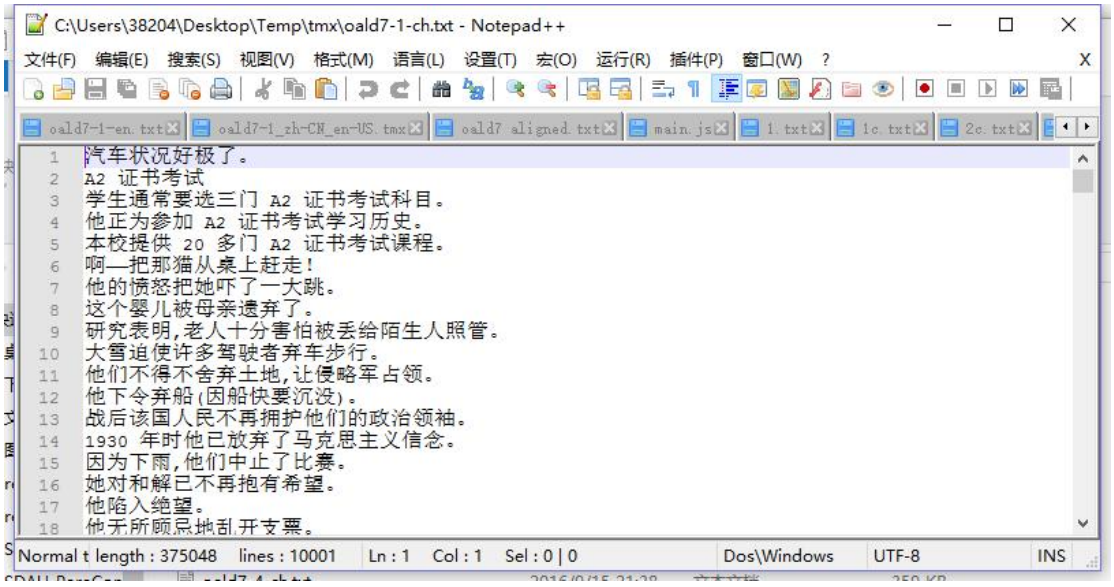
单文件对齐语料是指两种语言存放在同一 txt 文件里，每个句子一行。可以是英语在前也可以是汉语在前，如下图所示：



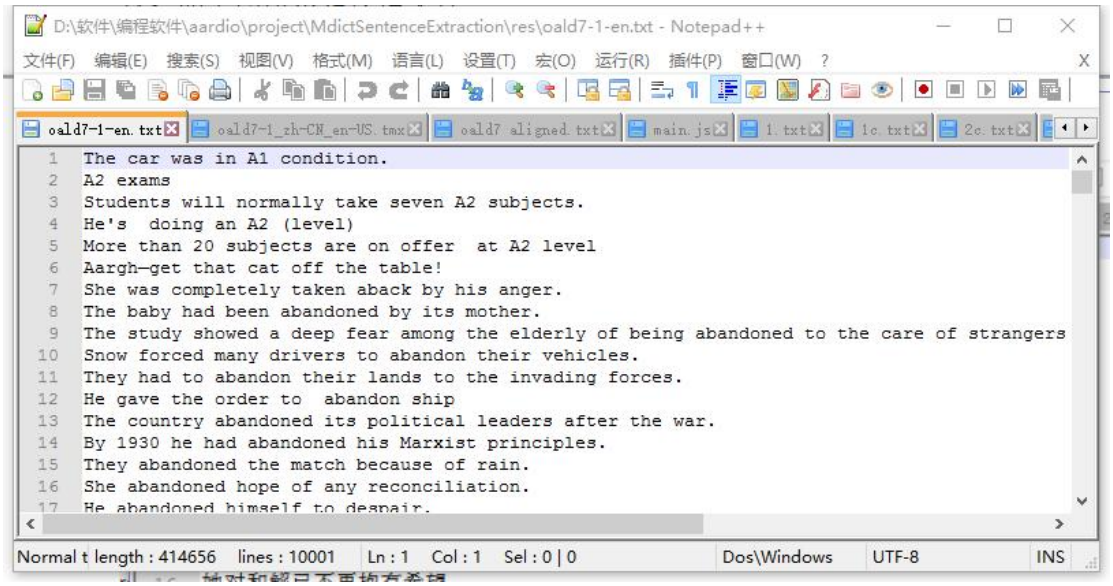
上图是汉语在前，下图是英语在前：



双文件对齐语料是指每个文件只存放一种语言。每行一句。英语和汉语对应句子所在行数应当一致。如下图所示是汉语文件：



下图是跟汉语文件对应的英语文件：

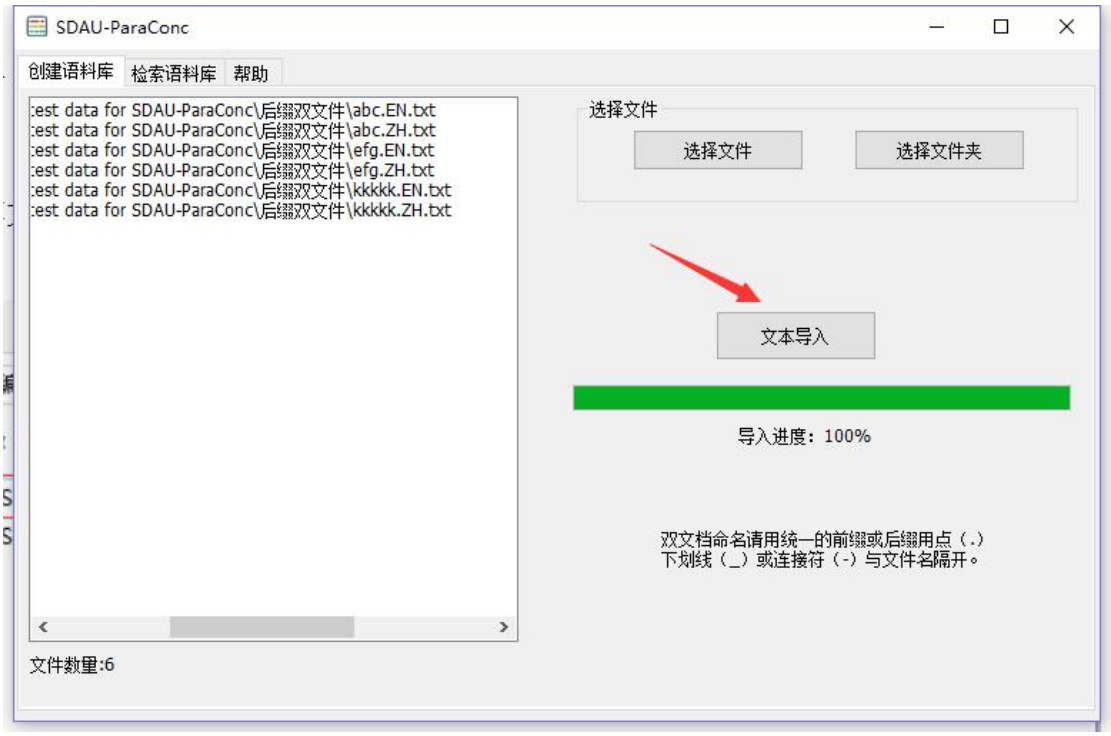


请注意：如果是双文件语料，英汉两个对应的**文件名应当一致**，且文件名应当有**统一**的语言识别标识（前缀或后缀），请用点（.）下划线（_）或连字符（-）将前缀或后缀与文件名隔开。
如：您可以将汉语文件命名为 abc.zh.txt 其对应的英文文件名应当为 abc.en.txt
语言标识可以是后缀，如上例，也可以是前缀：
如： 汉语文件名 ch-abc.txt, 英语文件名 en-abc.txt

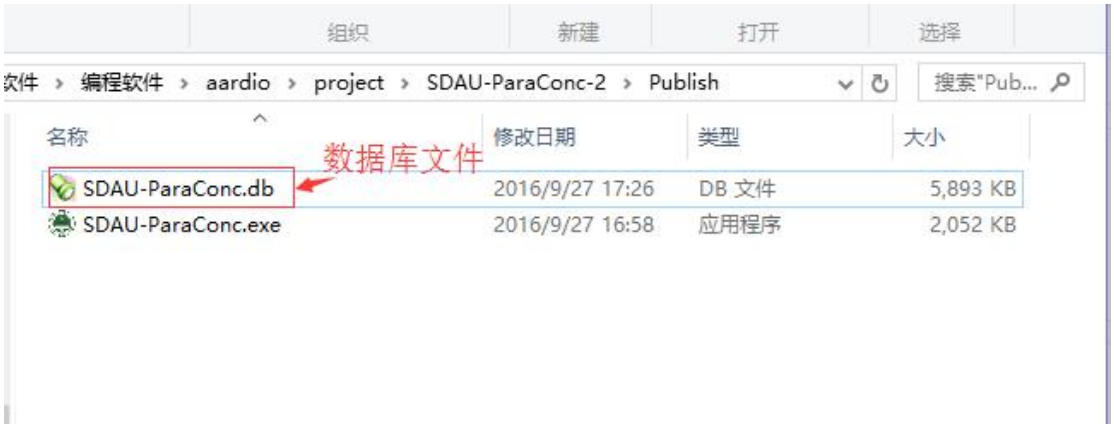
注意：如果您在检索时发现双文件语料没有对齐，请检查您的文件命名是否符合上述规则。

Tmx 文件支持英汉或汉英的对齐语料。如您需要检索其他语言的 tmx 文件，请联系作者。

5 点击“文本导入”按钮，软件会将语料导入 sqlite 数据库文件。



数据库文件名为 SDAU-ParaConc.db，存放在软件所在目录。



每次您点击“文本导入数据库”按钮，软件都会清空 SDAU-ParaConc.db 并重新写入新的数据。
如果您下次直接进入检索页面，软件默认检索上次生成的数据，也就是 SDAU-ParaConc.db 里的数据。

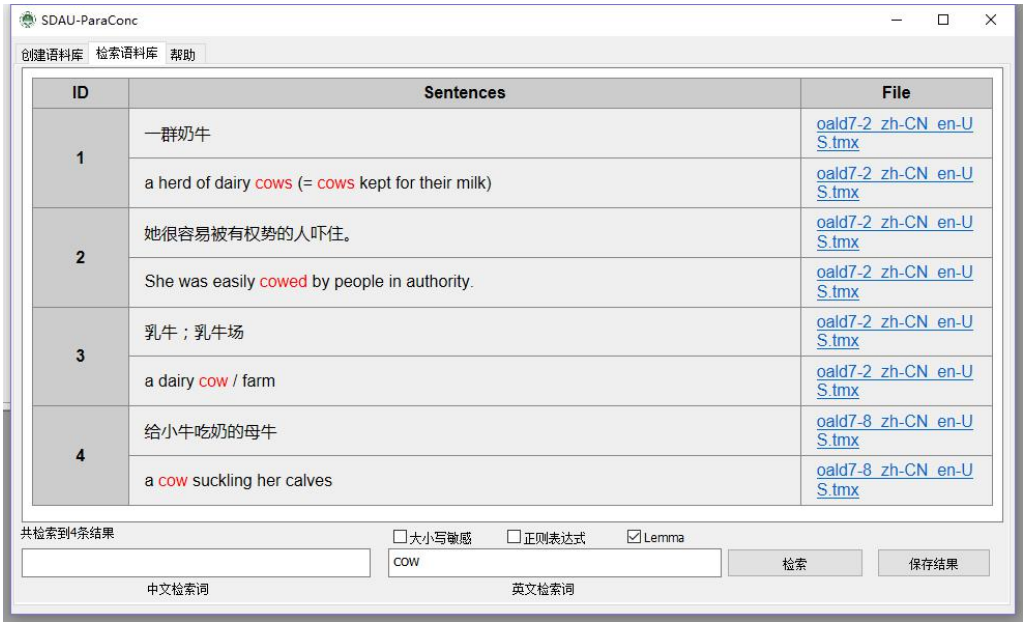
二、检索语料库

当您导入数据或选择现有数据后，就可以进行检索了。
输入您想要的检索的汉语和英语的检索词，点击“检索”按钮即可。

您可以同时输入汉语和英语检索词，也可以只输入任何一个。
检索词用红色标出。

英文检索词支持大小写敏感和 Perl 正则表达式。

您可以在 http://www.jb51.net/tools/zhengze.html 学到正则表达式。



勾选 Lemma 复选框会进行 lemmatized 检索，如输入 cow 会检索出 cow 的所有屈折变化
程序所在文件夹下的 lemmalist.db 是 lemmalist 数据库

点击最右列文件名，将会**尝试打开**该句所在文件。

点击保存结果按钮，软件将保存当前检索结果到.html 文件。您可以用任何一款浏览器打开。
如果勾选保存结果按钮上的 txt 格式，则会保存为 TXT 格式：每一行包含汉语文本、英语文本、汉语文件路径、英语文件路径四个部分，各部分之间用制表符（Tab）分开。

作者邮箱：gexiaoshuai89@foxmail.com
作者主页：http://gexiaoshuai.top

引用：葛晓帅. (2017). SDAU-ParaConc 0.0.3.01. 山东农业大学外国语学院.
Ge Xiaoshuai. (2017). SDAU-ParaConc 0.3.01. College of Foreign Languages, Shandong Agricultural University.

欢迎您的建议和意见。

Bug 修复和功能添加：
0.0.0.5 修复数据库查询限制为空时出错

0.0.0.6 修复读取文件第一行为空时无法自动判断单文件语言顺序问题

0.0.0.7 检索界面增加“保存结果”功能

0.0.0.8 修复单语双文件用前缀时出错问题

0.0.0.9 将检索结果字体改为 sans-serif 系列，默认 Arial 字体。

0.0.0.10 修复有时读取 ansi 编码文本乱码的问题

0.0.0.11 优化检索结果界面

0.0.1.0 增加 lemma 查询功能

0.0.1.0 修正正则表达式检索分支条件 bug

0.0.1.0 改进保存结果文件名自动完成功能

0.0.1.01 修正选择现存数据库后再次生成数据库不更新 bug

0.0.2.01 增加结果保存为 TXT 格式功能

0.0.2.02 修复导入 tmx 文件时有时乱码 bug

0.0.3.01
更换软件图标；
增加并简化测试数据；
自动判断语料类别，简化语料库创建步骤。
感谢许家金教授提出宝贵意见。