如何在 Déjà Vu X3 中自定义断句规则

在使用计算机辅助翻译软件的过程中，我们往往会遇到一些断句不当的情况，譬如，在源语言为英文的项目文件中，CAT 工具有可能会在“1.”或“Mr.”中的“.”处断句。那么，为了避免这种情况的出现，我们该如何解决呢？那当然是通过自定义断句规则就可以解决此类问题。今天让我们一起来学习如何在 Déjà Vu X3 中自定义断句规则：

Déjà Vu 中断句规则的工作原理：Déjà Vu 通过浏览文本，找到与既定断句规则匹配的文本时，便在此处将句子拆分为单独句段，从而实现自动拆分句段。然而，在执行拆分之前，Déjà Vu 会检查找到的文本是否符合设定的例外情况；如果符合，Déjà Vu 便不会拆分该句段，会继续浏览。若想自定义断句规则，您可以使用任何实际的字母，加上一些 Déjà Vu 识别的符号，而这些符号用于表示特殊字符或字符组。

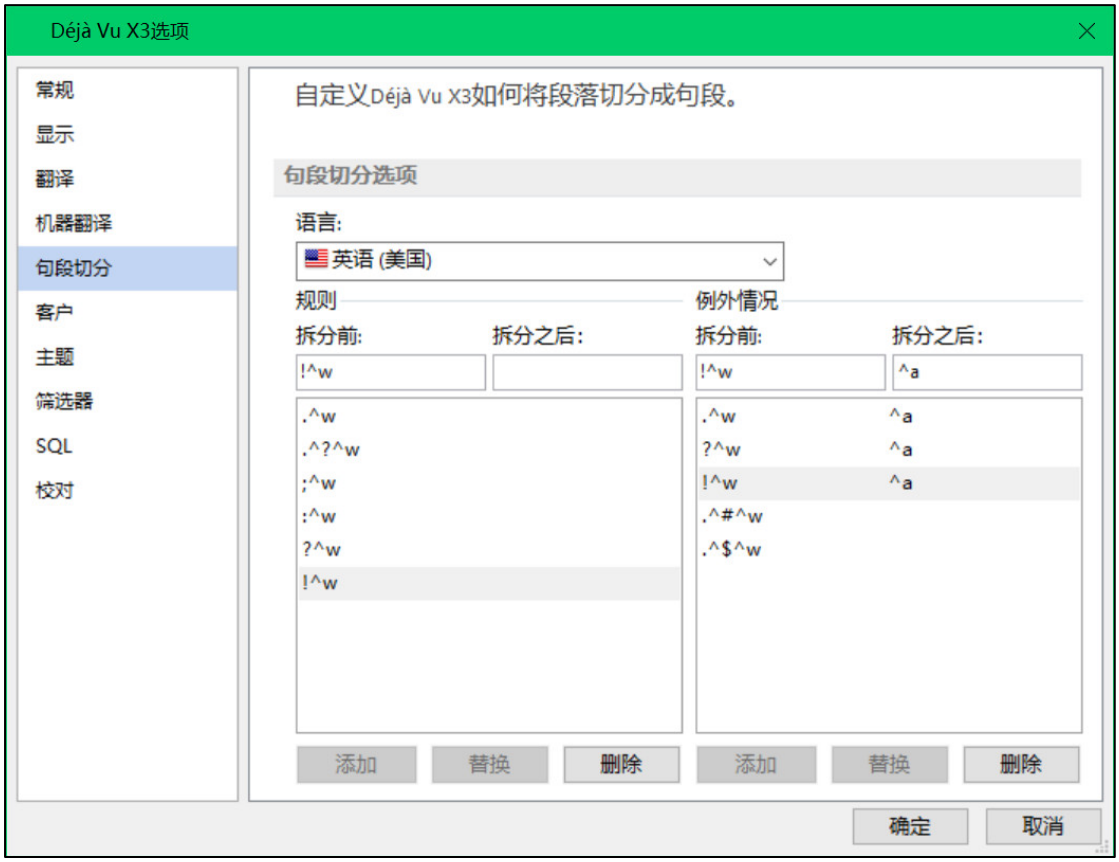
一、 符号：

符号	含义
^w	空格
^#	任意数字（1，2，3...）
^$	任意字母（大写、小写或任意大小写）
^a	小写字母
^A	大写字母
^?	任意字符
^^	插入符号字符（^）本身

二、 符号示例

1. 规则

以下是 Déjà Vu 在英语（美国）中使用的默认断句规则及其例外情况：



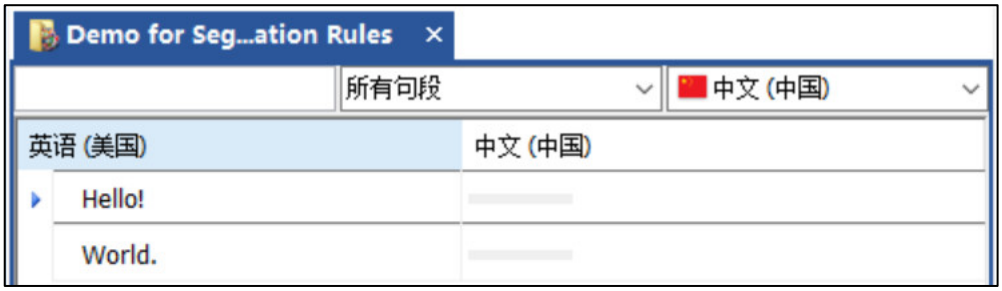
让我们看看第一条规则“!^w”。字符“！”代表它本身，就是一个感叹号。符号^w 代表一个空格。这意味着，每当 Déjà Vu 发现一个感叹号后跟一个空格时，就会在感叹号和空格之后拆分句段。因此，例文：

Hello! World.

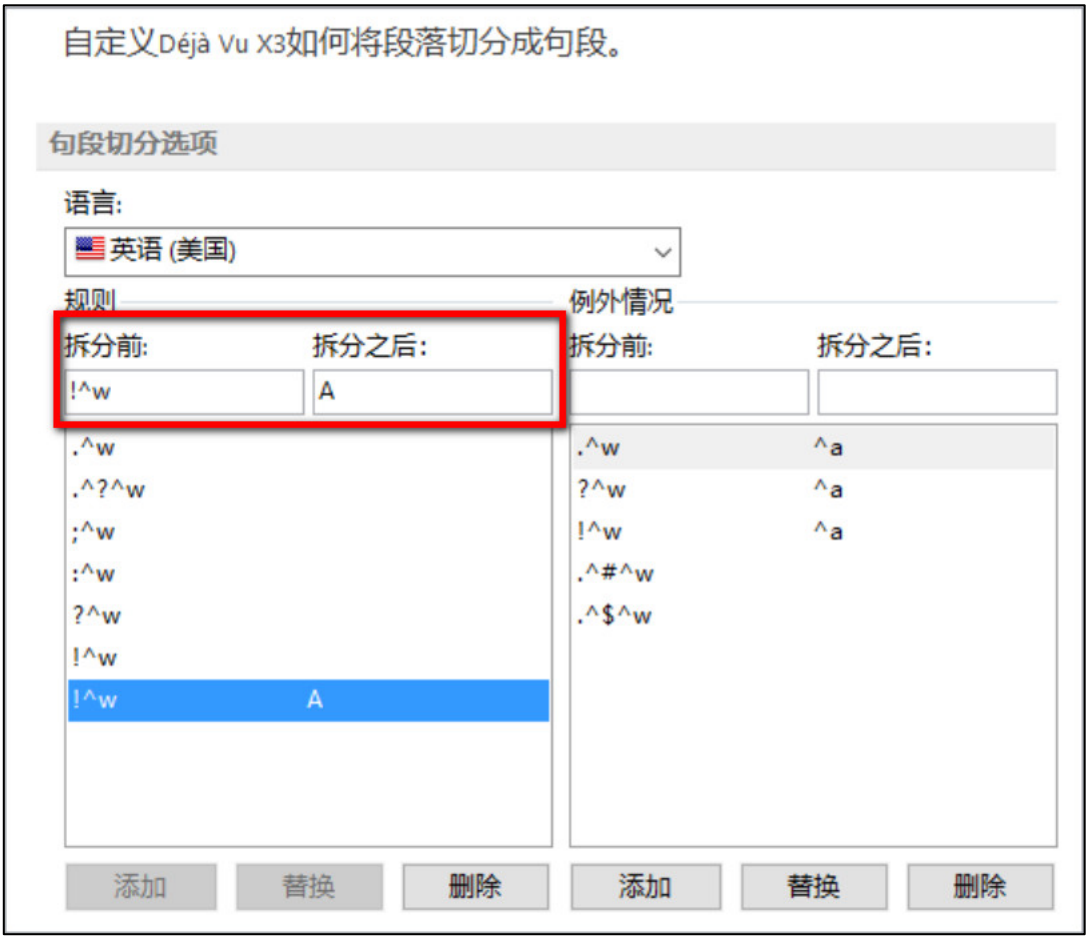
将拆分为：

Hello!

World.



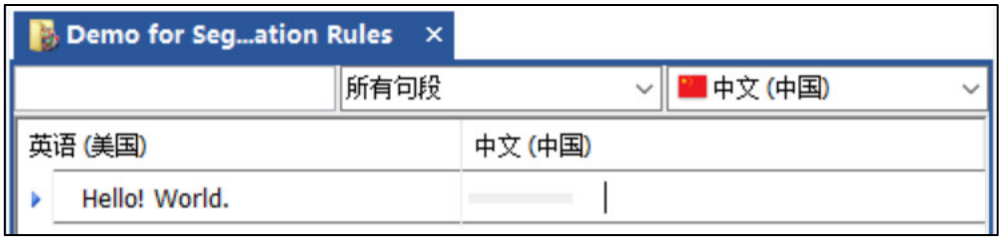
请注意，每个断句规则有两栏：“拆分点前”和“拆分点后”。在“拆分点前”一栏中，把 Déjà Vu 应该查找的内容放在拆分发生的地方之前，而在“拆分点后”一栏中，把 Déjà Vu 应该查找的内容放在拆分发生的地方之后。为了说明其工作原理，假设我们使用了这个断句规则：



在这种情况下，Déjà Vu 会在如下位置拆分文本：拆分点前有一个感叹号后跟一个空格，拆分点后有一个大写字母 A（请记住 A 和^A 不是一码事！）。根据这条规则：

Hello! World.

不会拆分。



但是：

Hello! A World.

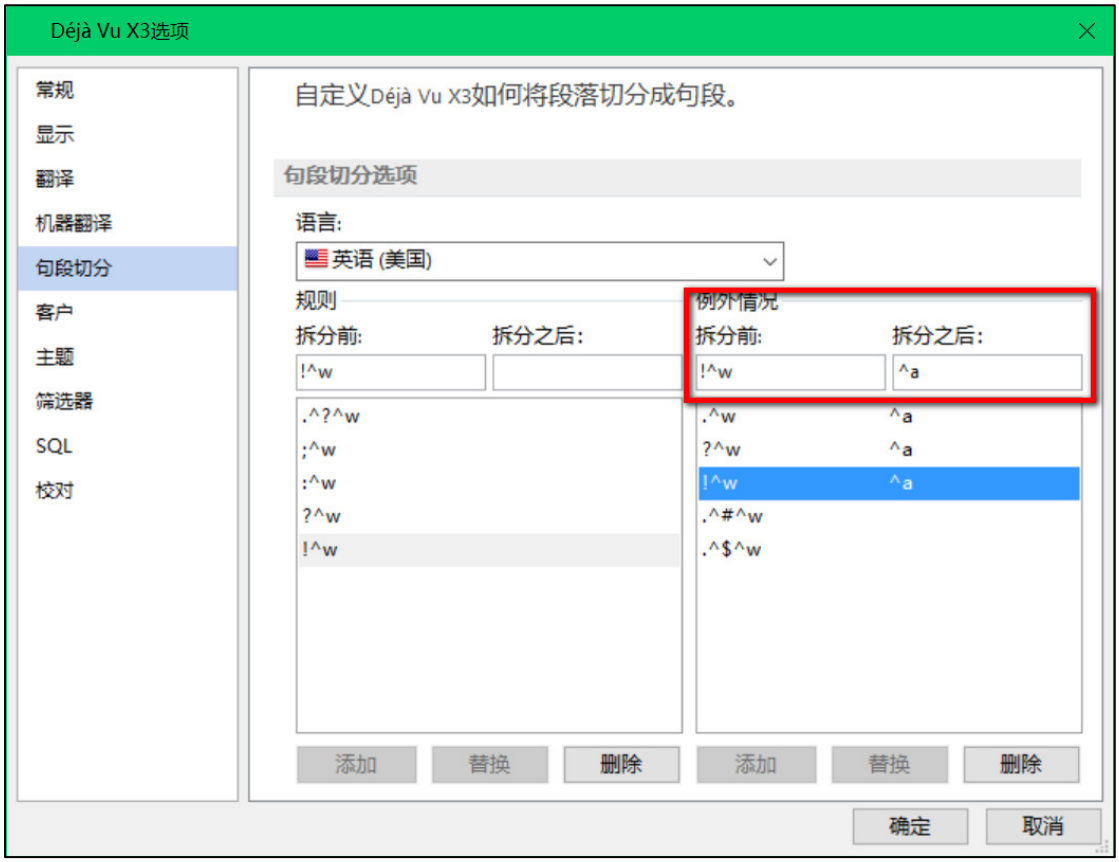
会被拆分。

将在大写字母 A 之前拆分句段。

2. 例外情况

例外情况在规则之后立即适用。如果 Déjà Vu 找到与其中一条规则匹配的文本，它将检查是否也与例外情况匹配。如果有匹配，Déjà Vu 将不会拆分，但若没有匹配，将继续拆分文本。

让我们看看第一条例外情况：



其含义是：在拆分点之前有一个感叹号后跟一个空格，并在拆分点之后有一个小写字母，Déjà Vu 必须破例，不予拆分。在没有例外情况的前提下，文本如下：

Use the big! service.

将在感叹号和空格之后拆分，但由于存在例外情况，将不会拆分。如果“service”一词以大写字母 S 开头，文本将被拆分。

三、 规则和例外情况的使用

1. 避免在“P.O.Box”处拆分句子

接下来一起看一个自定义的规则和例外情况的示例，让我们考虑一下如果文本包含“P.O. Box”会出现什么情况，例如：

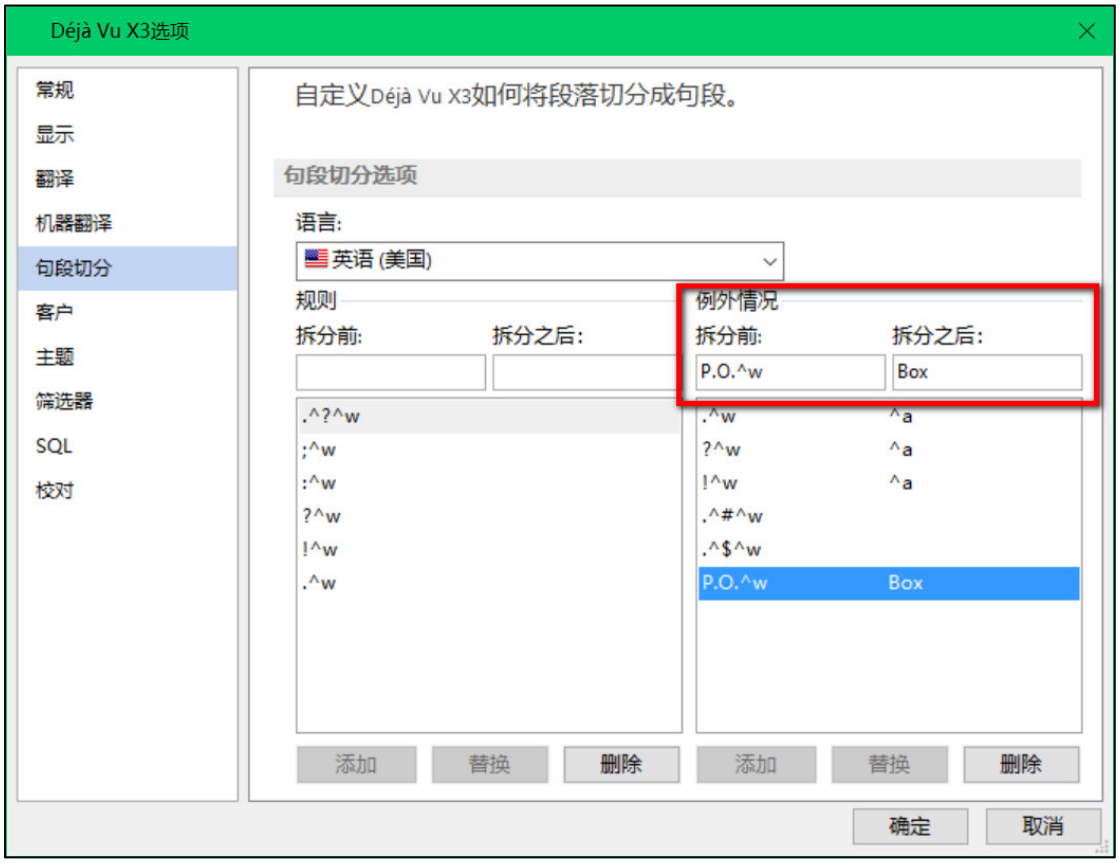
Acme can make deliveries to a P.O. Box as well as a physical address.

根据英语（美国）的默认规则，这将拆分为：

Acme can make deliveries to a P.O.

Box as well as a physical address.

之所以会出现这种情况，是因为 Déjà Vu 找到了一个句号后跟一个空格（“P.O.”之后），这意味着将在句号之后拆分，而其后面的文本（“Box”）与任何例外情况都不匹配。如何避免这种情况？考虑以下例外情况：



如果这一例外情况有效，当 Déjà Vu 确定“P.O.”之后但“Box”之前的位置是拆分的候选位置时，它会问：

“我将拆分的位置之前的文本是否包含字母“P.O.”后接空格？”

的确如此。它还会问：

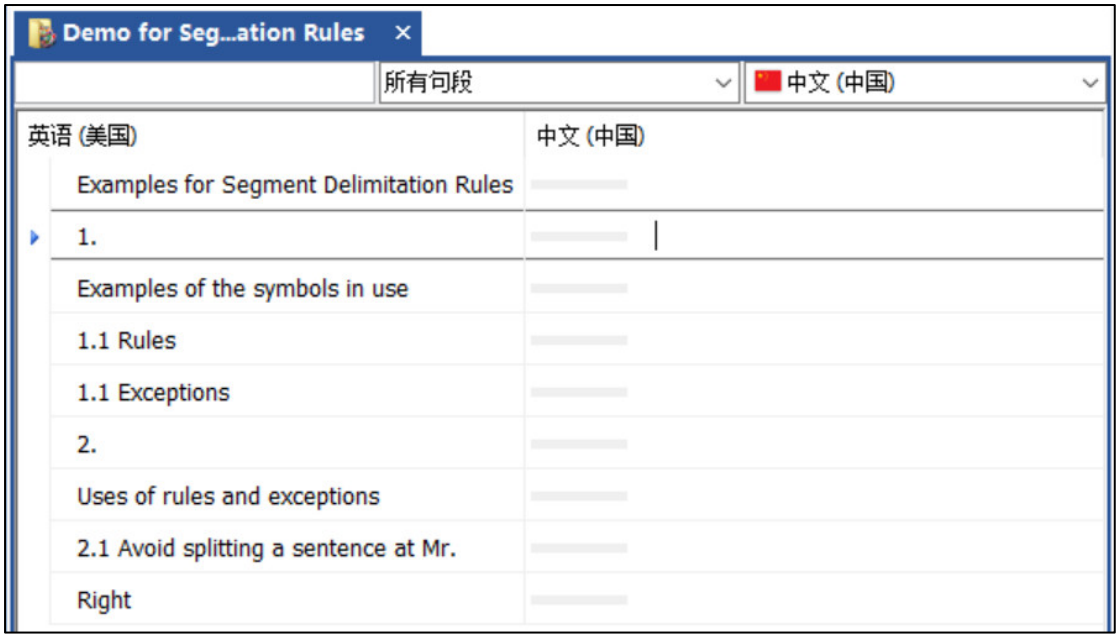
“我将拆分的位置后面的文本是否包含字母“Box”？”

是的，的确如此。因此，Déjà Vu 不会在这里拆分文本。

四、 案例实操

如果不想让 Déjà Vu X3 在“1.”或“Mr.”中的“.”处断句，我们该如何操作呢？

首先，我们先看看 Déjà Vu X3 默认断句规则对“1.”或“Mr.”的处理结果：



这样的话，我们需要手动对一些句段进行合并，如果这种情况在项目中出现频率较高，我们建议自定义断句规则后再重新导入文件，具体操作如下：

1. 梳理自定义断句规则例外情况的表达式

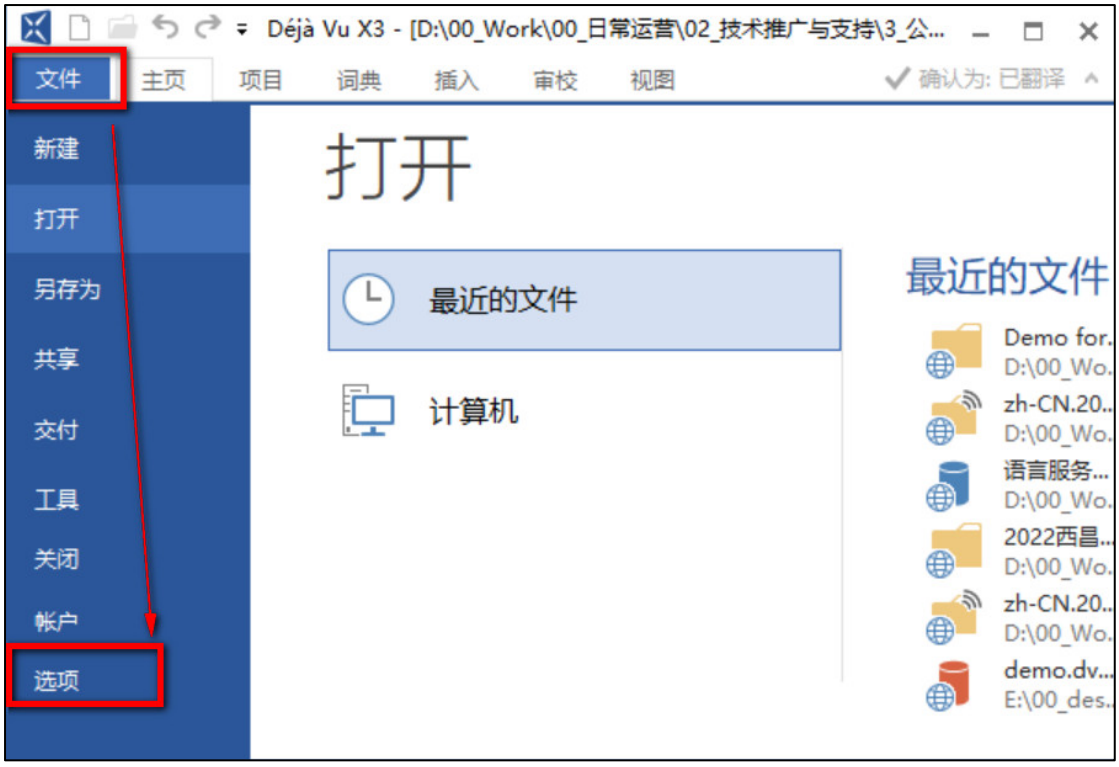
Déjà Vu X3 之所以会在“1. Examples of the symbols in use”、“2. Uses of rules and exceptions”和“Mr.”中的“.”处断句，是因为检测到默认断句规则“.^w（半角句点后跟空格）”，而没有检测到匹配的例外情况，所以我们需要在默认断句规则的前提下再自定义例外情况的表达式，经过观察，不难发现“1.”或“Mr.”断句之前的元素可用以下实际字母加上一些符号来表示：

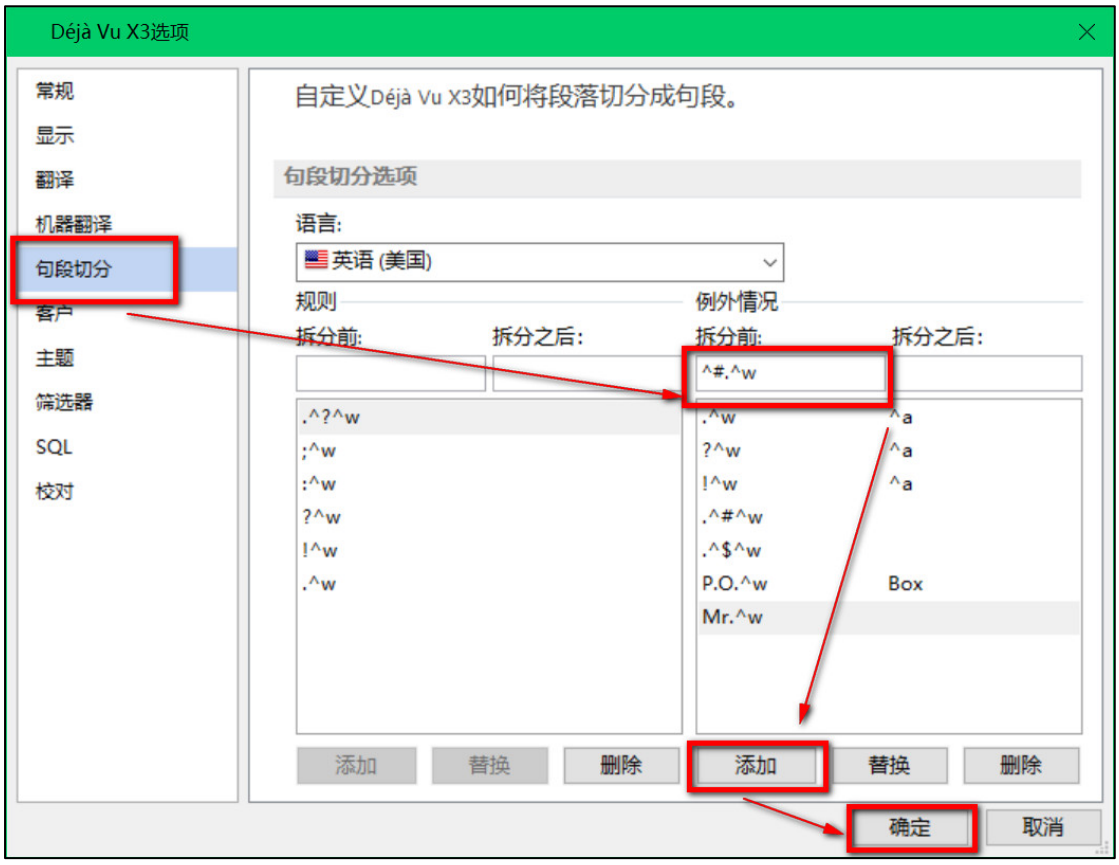
“1.”：数字+句点+空格，换成 Déjà Vu X3 能识别的规则就是：^#.^w

“Mr.”：Mr+句点+空格，换成 Déjà Vu X3 能识别的规则就是：Mr.^w

2. 添加自定义例外情况的表达式

1) 在打开项目文件的界面下，点击“文件”>“选项”，在弹出的窗口中，切换到“句段拆分”，并将例外情况的规则“^#.^w”和“Mr.^w”添加到“例外情况”列表中：





2) 添加完例外情况规则后，在“项目浏览器”下重新导入源文件即可。