关于高级汉语学习者误用语料库（满月语料库）的构建

张学博（东京外国语大学研究生院）

高级汉语学习者误用语料库（满月语料库）

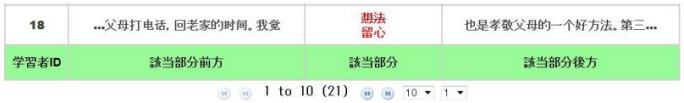
- 语料库特征

・最后将以网上公开的形式发布。

・搜索条件的多样化(包括误用的类型，学习者的背景资料等等)。

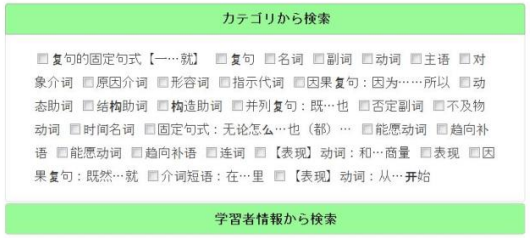
・搜索的结果中，不仅可以看到出现误用的句子，也可以阅览相关的原文。

- 样本



学習者ID	該当部分前方	該当部分	該当部分後方
18	…父母打电话，回老家的时间，我觉	想法 留心	也是孝敬父母的一个好方法，第三…

1 to 10 (21)



カテゴリから検索

□ 复句的固定句式【一…就】 □ 复句 □ 名词 □ 副词 □ 动词 □ 主语 □ 对象介词 □ 原因介词 □ 形容词 □ 指示代词 □ 因果复句：因为……所以 □ 动态助词 □ 结构助词 □ 构造助词 □ 并列复句：既…也 □ 否定副词 □ 不及物动词 □ 时间名词 □ 固定句式：无论怎么…也（都）… □ 能愿动词 □ 趋向补语 □ 能愿动词 □ 趋向补语 □ 连词 □ 【表现】动词：和…商量 □ 表现 □ 因果复句：既然…就 □ 介词短语：在…里 □ 【表现】动词：从…开始

学習者情報から検索

学習者ID	該当部分前方	該当部分	該当部分後方
18	…父母打电话，回老家的时间，我觉	想法 留心	也是孝敬父母的一个好方法，第三…

1 to 10 (21)

カテゴリから検索

学習者情報から検索

TOEFL（iBT）：0点 - 120点

学習期間：0年 - 20年

海外滞在経験：□あり □なし

性別：□男 □女

年齢：□１０代 □２０代 □それ以上

国籍：□日本 □中国 □中国

母語：□日本語 □中国語 □英語 □その他

教育言語：□日本語 □中国語 □英語 □その他

中国語学習者中国語作文誤用コーパス検索サイト

サイトについて

本サイトは、中国語学習者コーパスのデータを検索できるWebアプリケーションです。コーパスデータは、東京外国語大学・中国語専攻、望月圭子研究室、佐野洋研究室、中垂敏、福田翔両 氏らの協力のもと収集されました。検索が可能なデータ件数は、平成25年10月○日時点で○○件です。

誤用文字列：⇒ 修正文字列：検索 リセット

タイプ検索：□置換 □削除 □追加

1 to 10 (21)

学習者ID	該当部分前方	該当部分	該当部分後方
18	…东京。东京的生活费用比我老家太	贵 高	为减轻那个负担，我现在打工挣钱…
18	…我老家贵，所以父母的负担很大，	为 为了	减轻那个负担，我现在打工挣钱，…
18	…贵，所以父母的负担很大，为减轻	那个 他们	负担，我现在打工挣钱，我挣的钱…
18	…自己打工挣一些钱，我们的负担	减小 减少	了，"第二个是时间的方面，上大…
18	…每天很忙，没有时间，所以我没有	对 给	父母打电话，或者去老家的时间，…
18	…间，所以我没有对父母打电话，	或者 也	去老家的时间，但是，毕业大学，…
18	…间，所以我没有对父母打电话或者	去 回	老家的时间，但是，毕业大学，弄…
18	…以，我现在改变了自己想法，无论	如何 怎么	忙，也抽对父母打电话回老家的时…
18	…变了自己想法，无论如何忙，也抽	对 给	父母打电话，回老家的时间，我觉…
18	…父母打电话，回老家的时间，我觉	想法 留心	也是孝敬父母的一个好方法，第三…

学習者ID | 該当部分前方 | 該当部分 | 該当部分後方

作业流程

前期资料整理

- 通过收集承诺书来获得汉语学习者的相关个人信息。
- 收集汉语学习者的作文
- 设计标签列表

构建语料库

- 修改作文
- 对误用之处添加标签
- 将添加完标签后的作文转换为XML网页文件

前期资料整理：收集承诺书·获得学习者信息

- 为了将来能够在网上公开学生所写的作文，所以在收集作文之前必须要请学生签署“承诺书”。

- 汉语学习者的个人信息

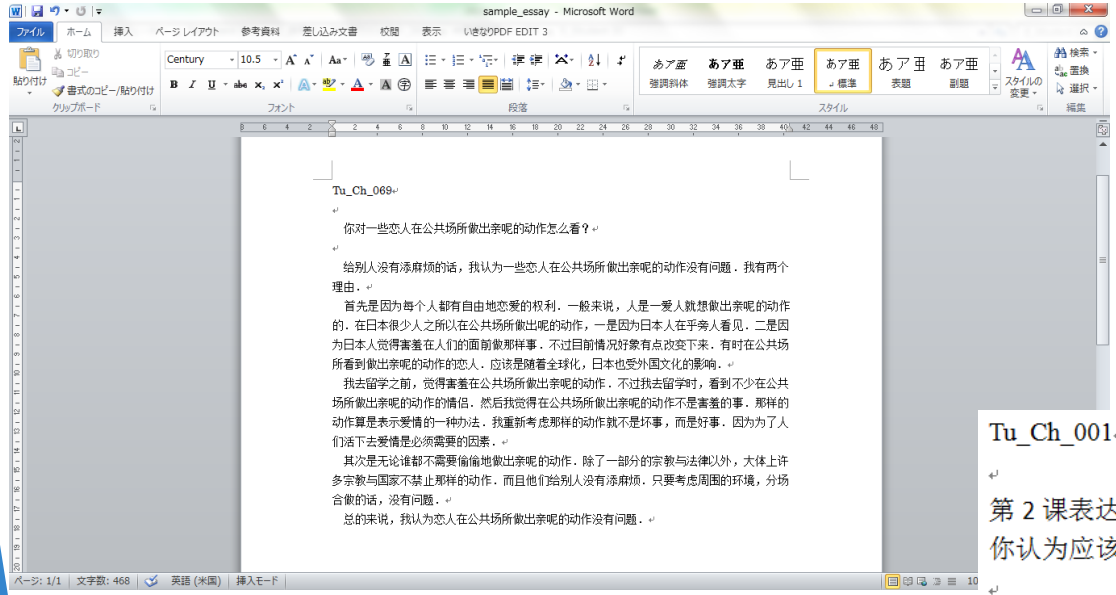
・国籍　・母语　・教育中所用语言　・汉语学习的时间　・学习汉语所在机构
・是否有留学经验(时间，机构，所在地)　・在家庭中所使用的语言
・和朋友交谈时所用语言　・小学教育中所用语言　・中学教育中所用语言
・高中教育中所用语言　・英语检定考试成绩　・TOEFL(iBT)　・TOEIC
・IELTS(academic)・IELTS(general)　・汉语检定考试成绩
・HSK(笔试)　・HSK(口试)

前期资料整理：收集作文
（目前为止收集到的作文情况）

- 对象：东京外国语大学汉语系2年级-4年级
- 人数：81名（2年级：38人　3年级：26人　4年级：17人）
- 数量：大约500篇（2013-2014学年）
- 字数：大约102,000字
- 收集方式：通过东外大所开发的Moodle系统（a type of software e-learning platform, a Learning Management System / Virtual Learning Environment）来实现学生提交作文和研究小组返还作文的流程。

学生所提交的作文样本

原　稿



修改后

Tu_Ch_001

第 2 课表达练习 3-3

你认为应该怎样孝敬父母？

我认为孝敬父母时，~~我~~ 要表示三个方面的感谢。第一个方面 是 ~~，~~ 关于 金钱方面的感谢。我老家在群马，不过现在我住在东京。东京的生活费用比我老家 ~~太~~ 贵，所以父母的负担很大。为 了 减轻 那个 他们的 负担，我现在打工挣钱。我挣的钱还没 不 那么多，不过我妈妈常常说，“ ~~，~~ 因为 你现在自己打工挣 了 一些钱， 所以 我们的负担减 ~~少~~ 少 了。”第二个是 关于 时间 的 方面 的感谢 。上大学以后，我的生活很充实，每天很忙，没有时间。所以我 既 没有 ~~对~~ 给 父母打电话，~~或者~~ 也 没有 回 ~~去~~ 老家的时间。

コメント [潘55]: 【删除】【主语】

コメント [S56]: 动词

コメント [S57]: 介词（对象）

コメント [m58]: 【替换】高很多【程度补语】

コメント [S59]: 介词 （原因）

コメント [潘60]: 【替换】他们【指示代词】

コメント [潘61]: 【添加】的【结构助词】

コメント [S62]: 代词

前期资料整理：标签列表

- 关于修改方法的标签

「删除」(removal)、「添加」(addition)、「替换」(replacement)、
「移动」 (transfer)（删除＋添加）

- 关于语法范畴的标签

参考图书：张斌 (2007)《新编现代汉语》复旦大学出版社.

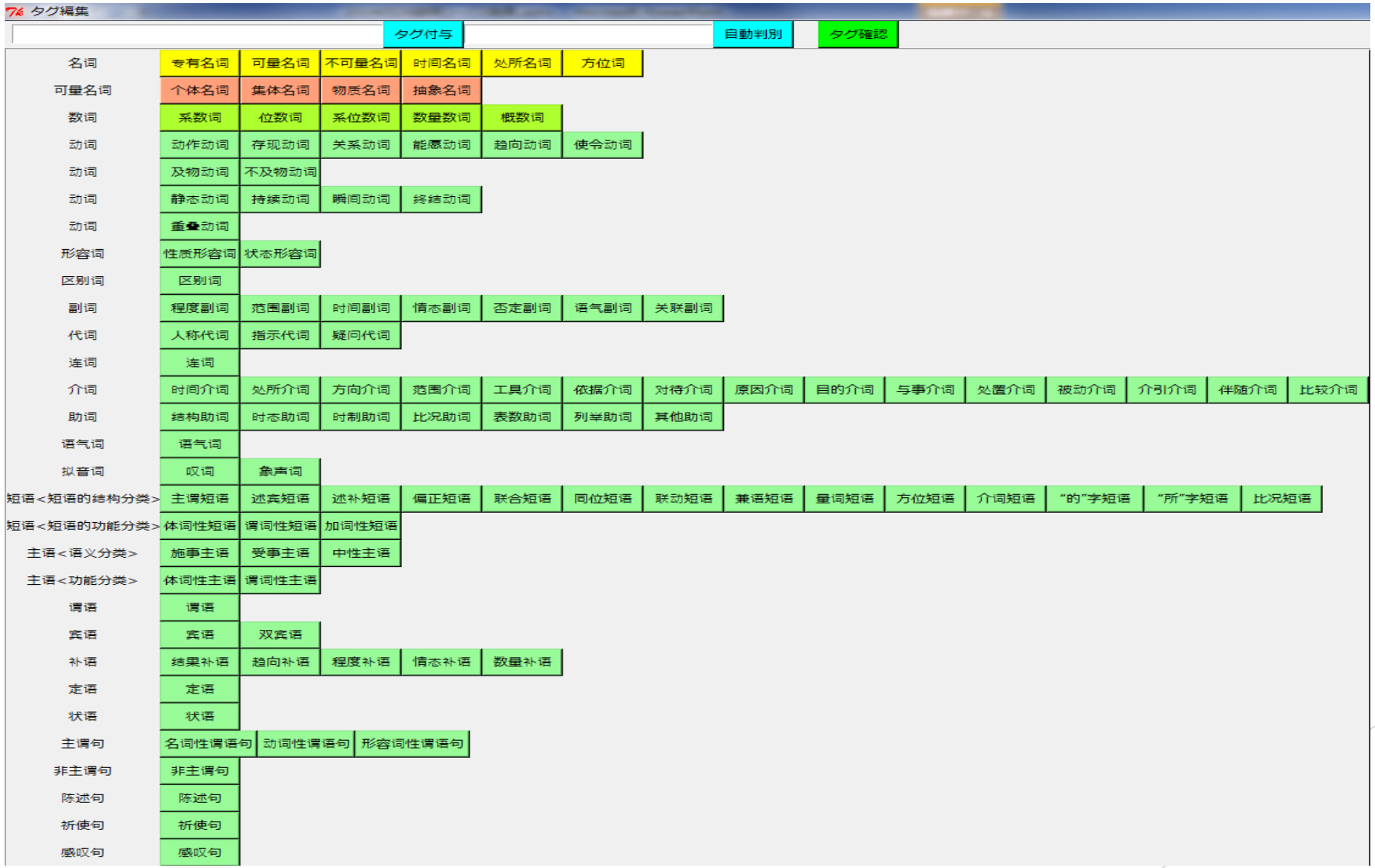
齐沪扬 (2012)《现代汉语》商务印书社.

齐沪扬 (2010)《对外汉语教学语法》复旦大学出版社.

→主要是根据以上三本来构建语法标签列表。

但并不是完全依照其理论，研究小组在添加标签的过程当中，会进行适当程度的简化或者细化。

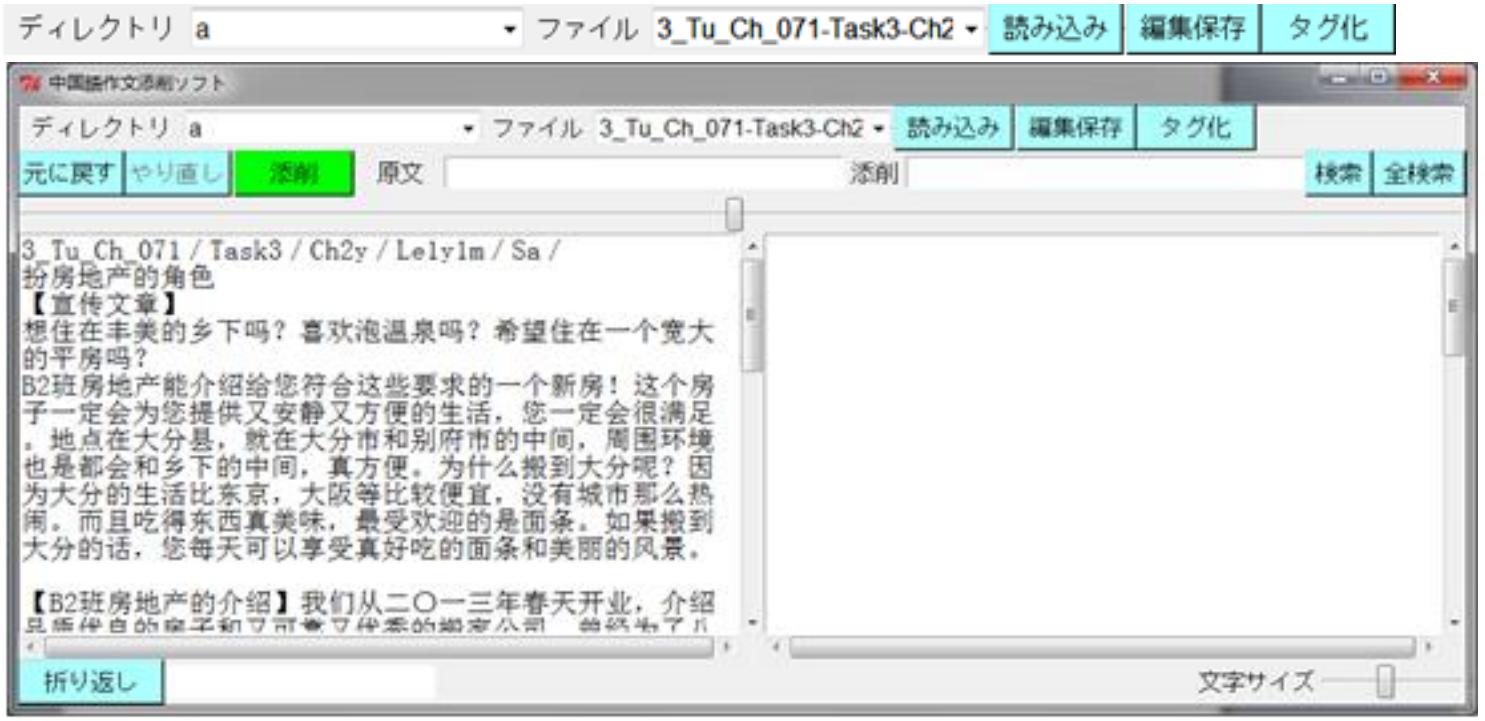
标签列表样本


タグ編集
タグ付与
自動判別
タグ確認
名词
专有名词
可量名词
不可量名词
时间名词
处所名词
方位词
可量名词
个体名词
集体名词
物质名词
抽象名词
数词
系数词
位数词
系位数词
数量数词
概数词
动词
动作动词
存现动词
关系动词
能愿动词
趋向动词
使令动词
动词
及物动词
不及物动词
动词
静态动词
持续动词
瞬间动词
终结动词
动词
重叠动词
形容词
性质形容词
状态形容词
区别词
区别词
副词
程度副词
范围副词
时间副词
情态副词
否定副词
语气副词
关联副词
代词
人称代词
指示代词
疑问代词
连词
连词
介词
时间介词
处所介词
方向介词
范围介词
工具介词
依据介词
对待介词
原因介词
目的介词
与事介词
处置介词
被动介词
介引介词
伴随介词
比较介词
助词
结构助词
时态助词
时制助词
比况助词
表数助词
列举助词
其他助词
语气词
语气词
拟音词
叹词
象声词
短语<短语的结构分类>
主谓短语
述宾短语
述补短语
偏正短语
联合短语
同位短语
联动短语
兼语短语
量词短语
方位短语
介词短语
“的”字短语
“所”字短语
比况短语
短语<短语的功能分类>
体词性短语
谓词性短语
加词性短语
主语<语义分类>
施事主语
受事主语
中性主语
主语<功能分类>
体词性主语
谓词性主语
谓语
谓语
宾语
宾语
双宾语
补语
结果补语
趋向补语
程度补语
情态补语
数量补语
定语
定语
状语
状语
主谓句
名词性谓语句
动词性谓语句
形容词性谓语句
非主谓句
非主谓句
陈述句
陈述句
祈使句
祈使句
感叹句
感叹句

构建语料库：修改作文

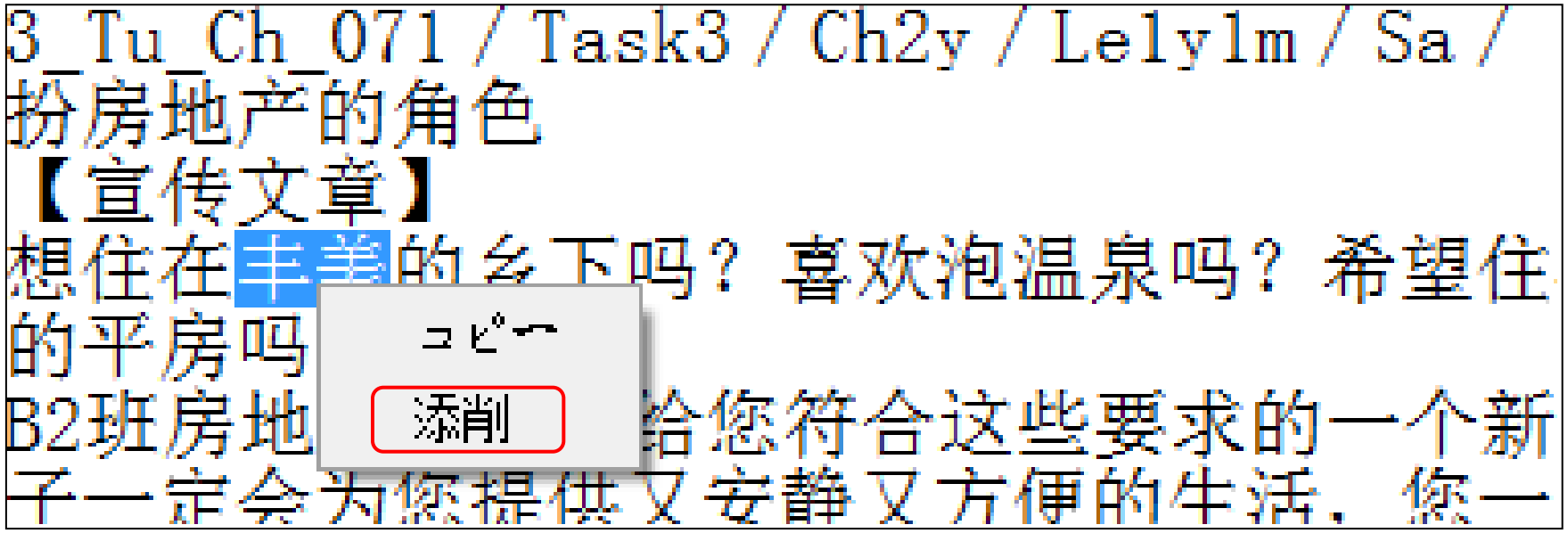
中国語誤用コーパスタグ付与ソフト（東京外大望月圭子科研専用）

- TNR_ChineseWritingCorrection.exe



- 本软件能实现的修改方法有四种：替换，删除，添加，移动。

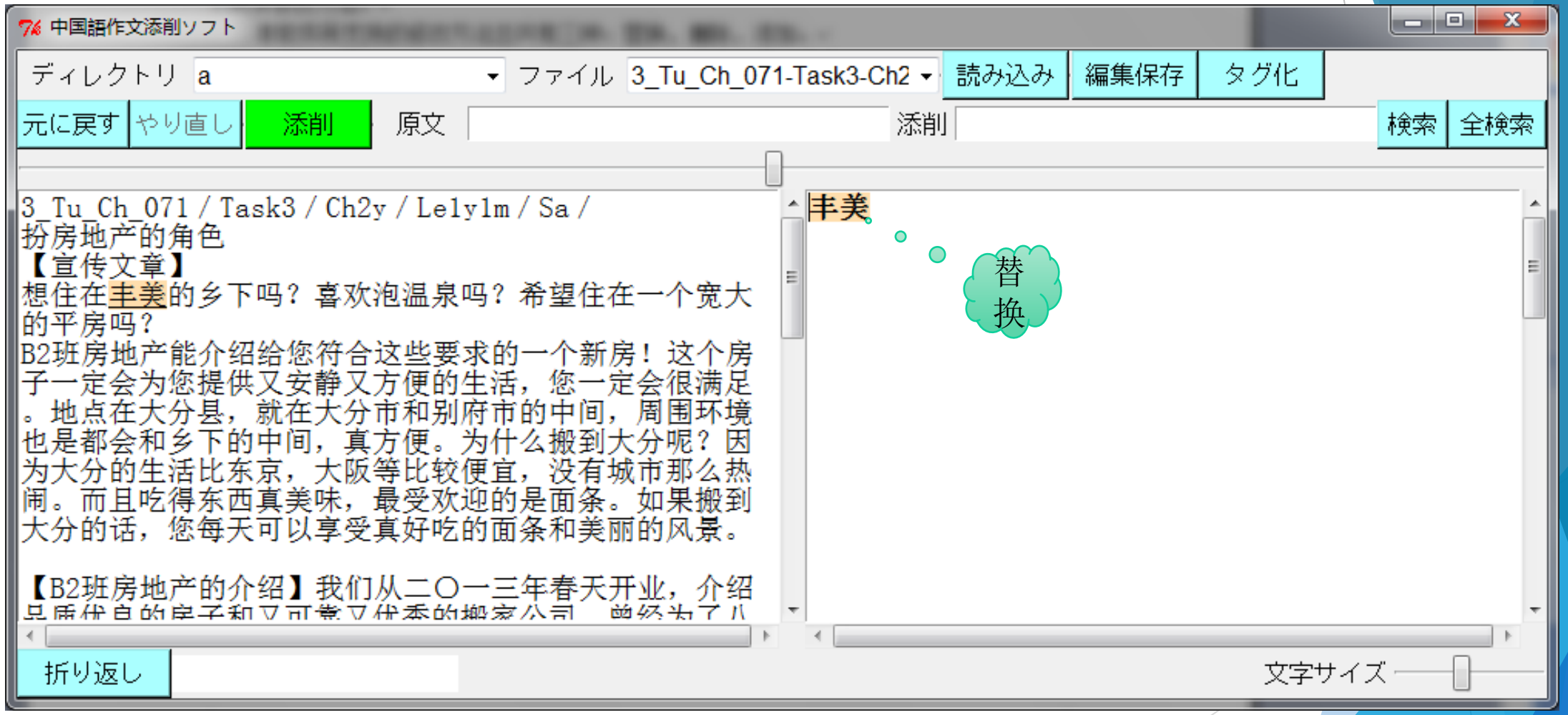
替换_1



选中误用处→右键→ 选择「添削」

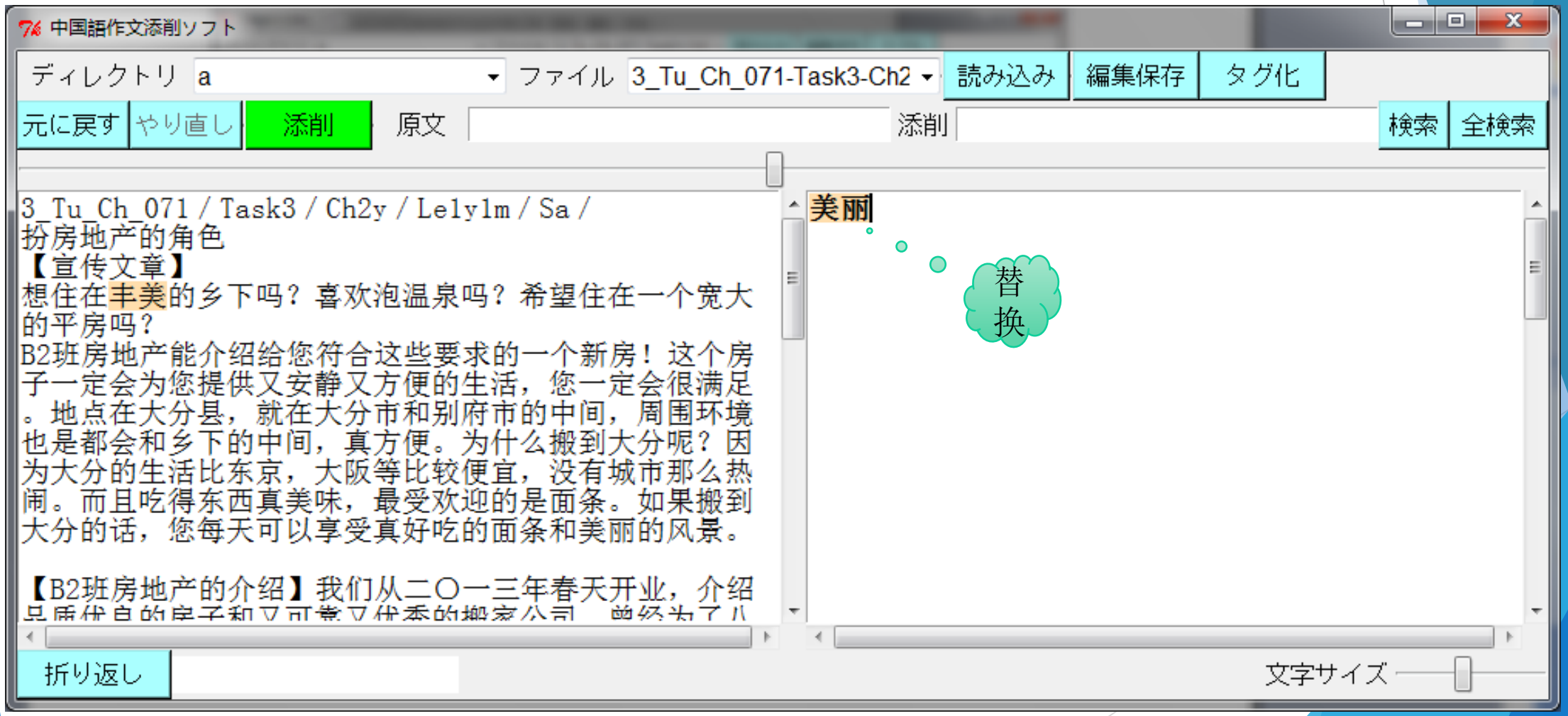
替换_2

界面的右半部分会自动出现刚才所选中的语句（〈丰美〉）。



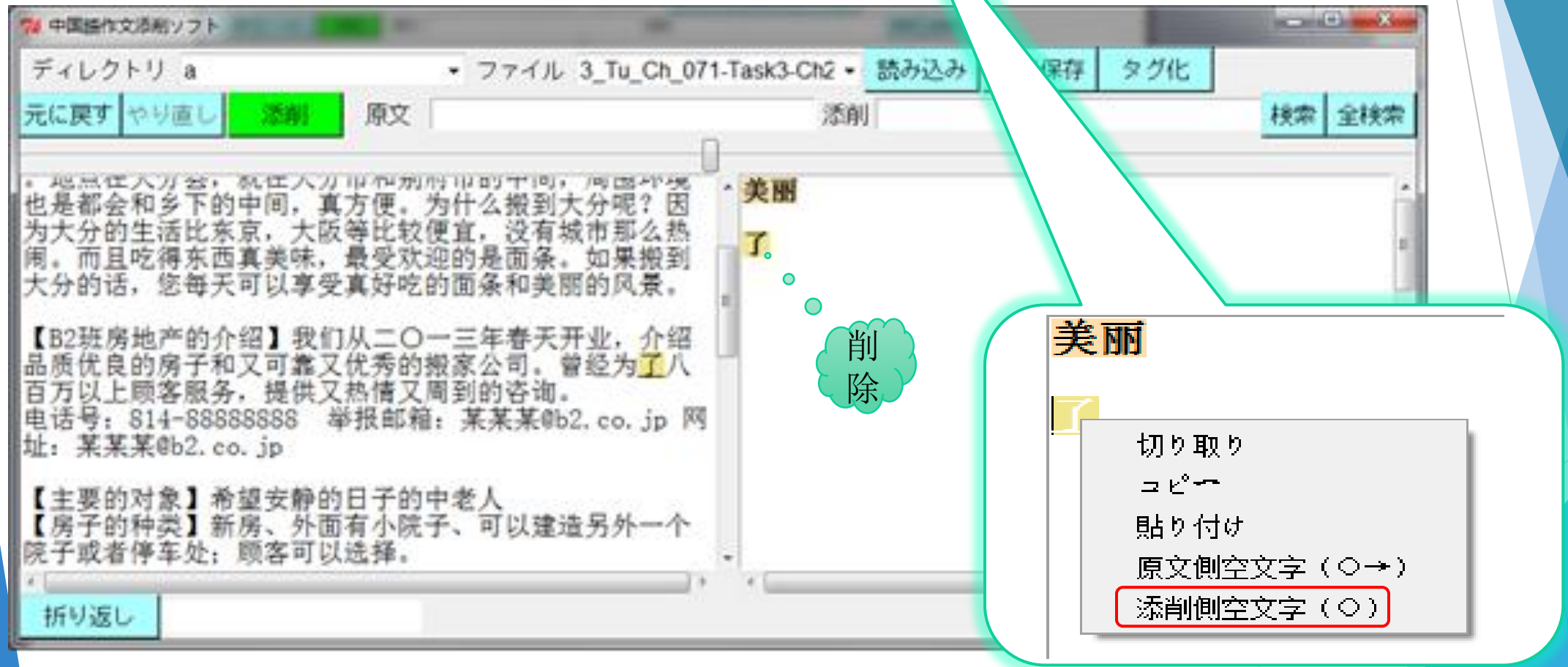
替换_3

直接删除界面有半部分的〈丰美〉，再直接输入适当的语句（〈美丽〉）即可。



删除_1

选中误用处→右键→ 选择「添削側空文字」



删除_2

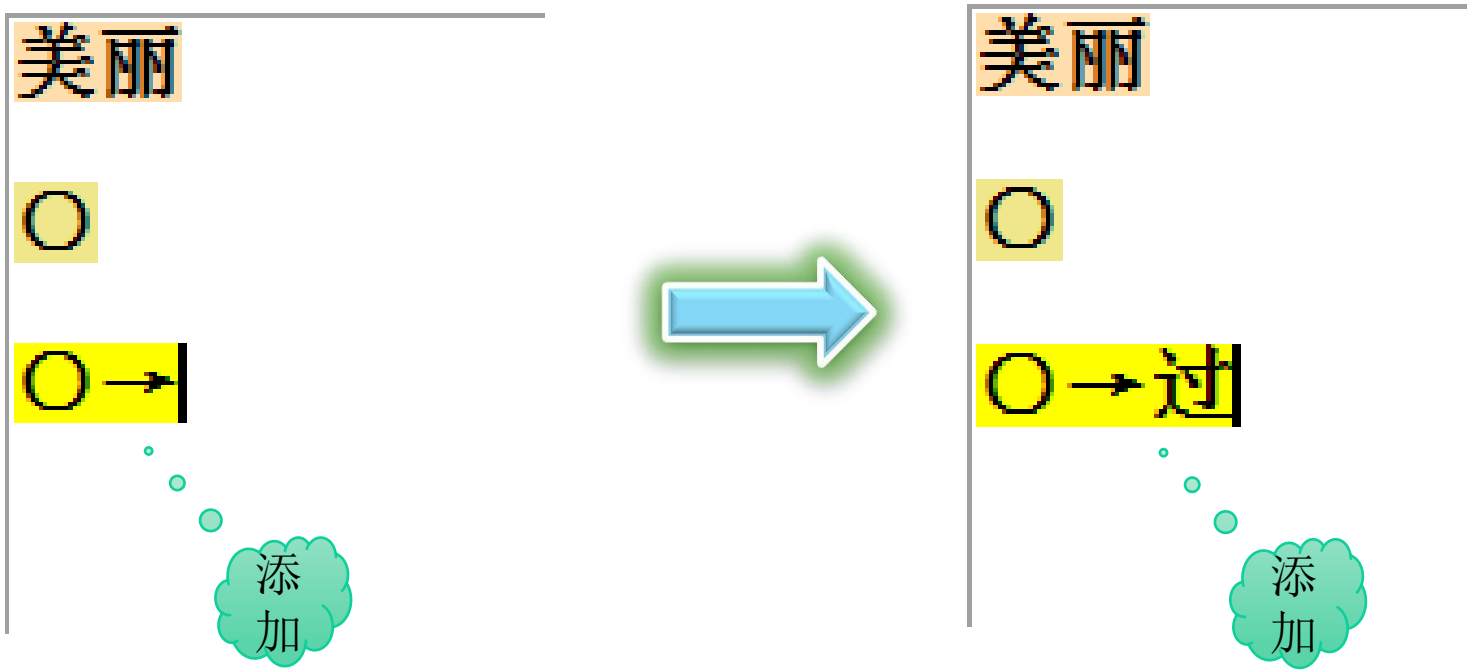
界面右半部分自动出现「〇」后，即完成删除的工作。

添加_1

选中误用处→右键→ 选择「原文側空文字（〇→）」

添加_2

右半部分界面会自动出现「○→」，然后再在「→」的后面直接输入需要添加的成分即可。

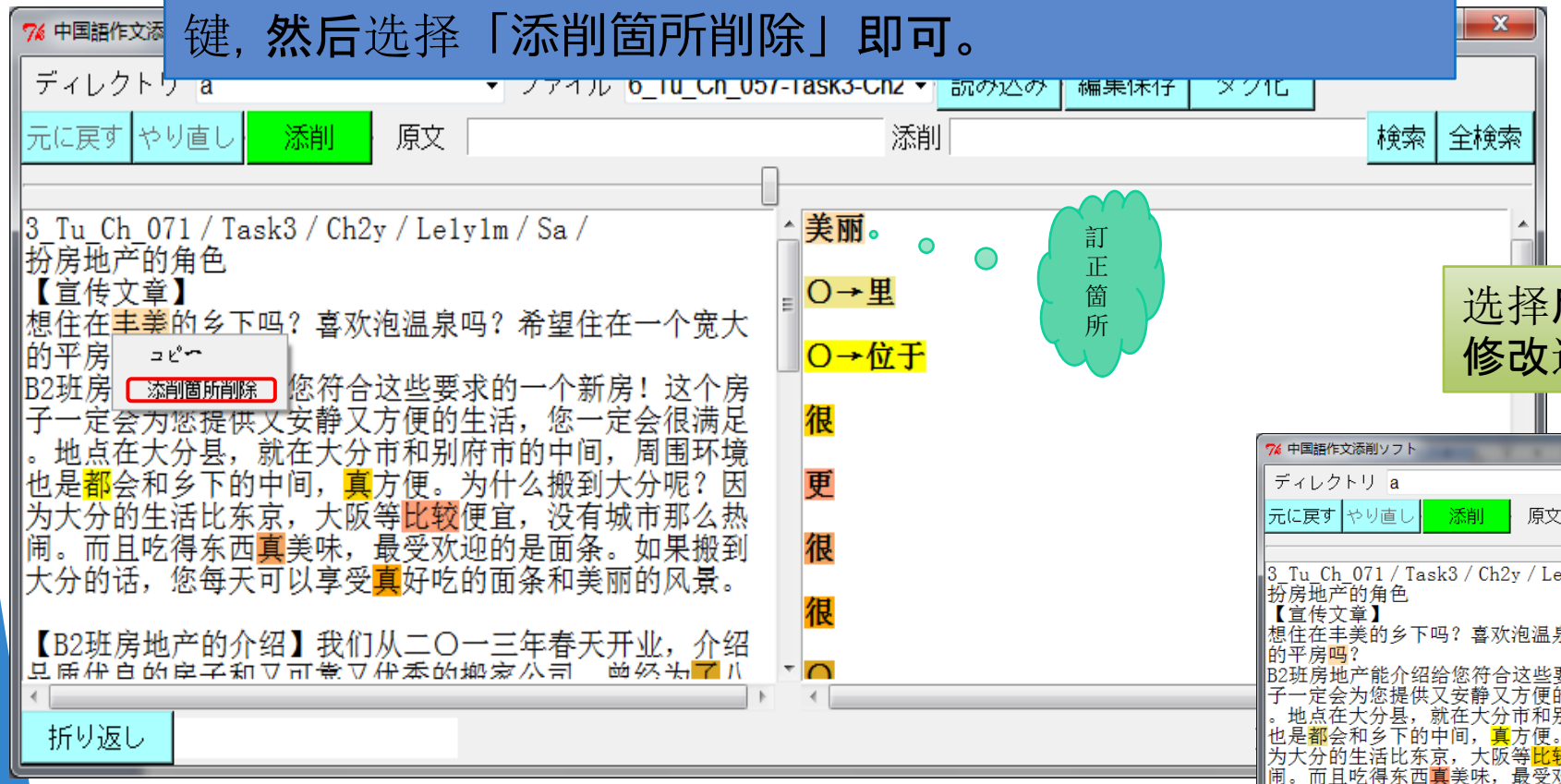


移动(删除＋添加)

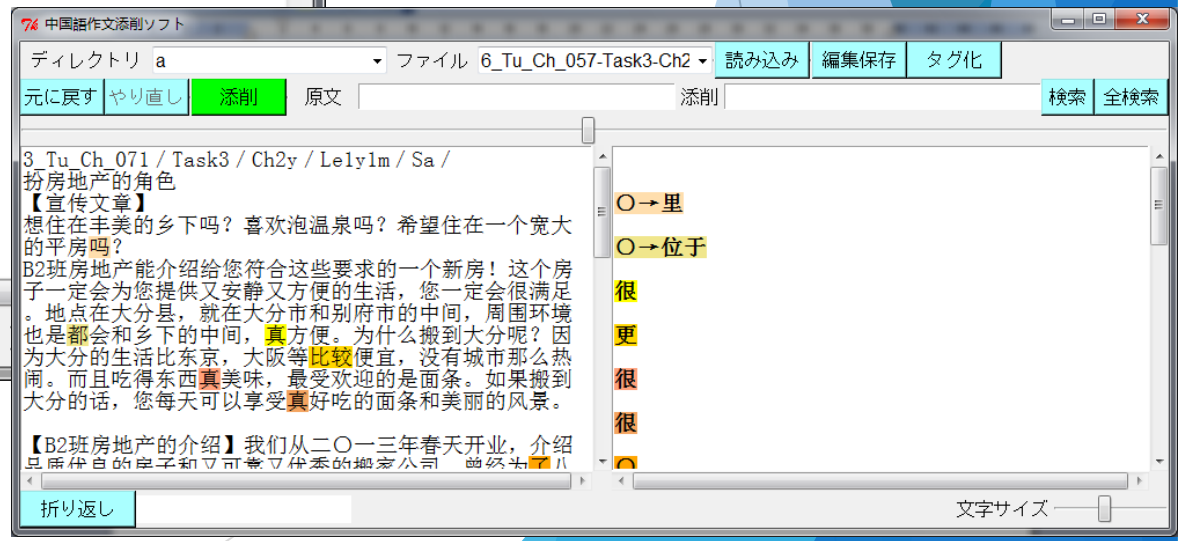
- 关于「移动」的操作仿佛，本软件无法直接实现，所以采用「删除」→「添加」的方法来完成。
- 「移动」主要针对于语序的误用问题。

关于删除之前修改过的误用处的方法

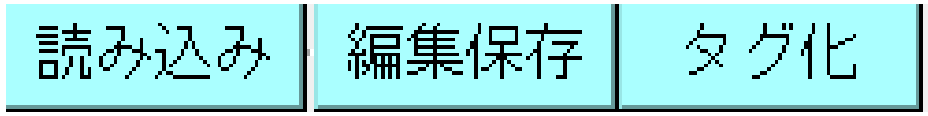
在界面左半部分需要删除的地方（〈丰美〉）的前面，右键，然后选择「添削箇所削除」即可。



选择后，即可看到界面右半部分之前修改过的地方自动消失。



保存修改完的作文

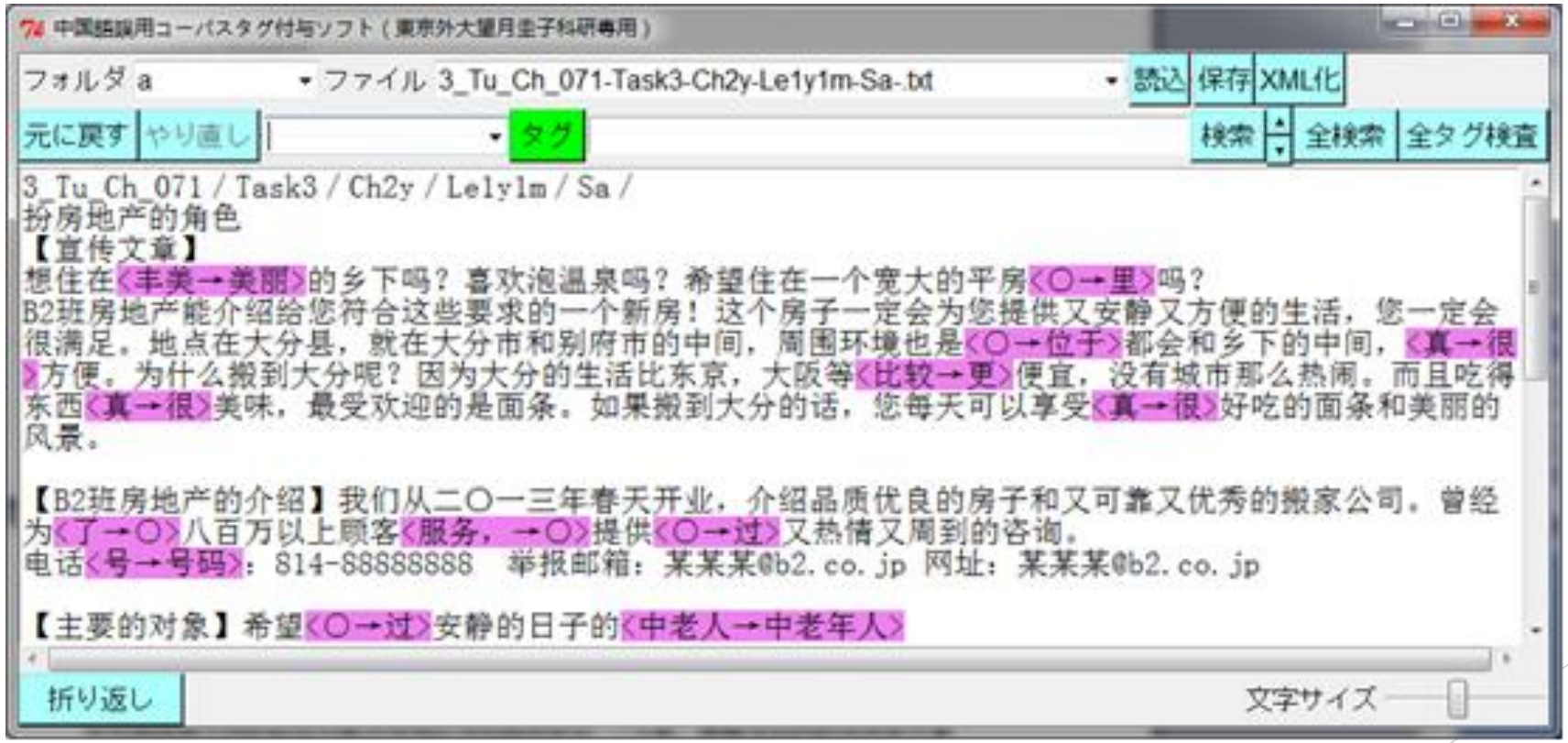


- 修改完成后，点击「编辑保存」即可保存完毕 。
- 之后在点击「タグ化」即可开始下一步添加标签的工作。

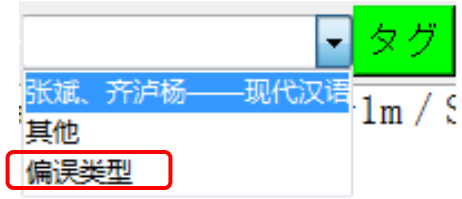
构建语料库：添加标签

- TNR_ChineseErrorCorpusTagger3.0_notag.exe 。

首先读取修改好的作文。读取后的界面如下：

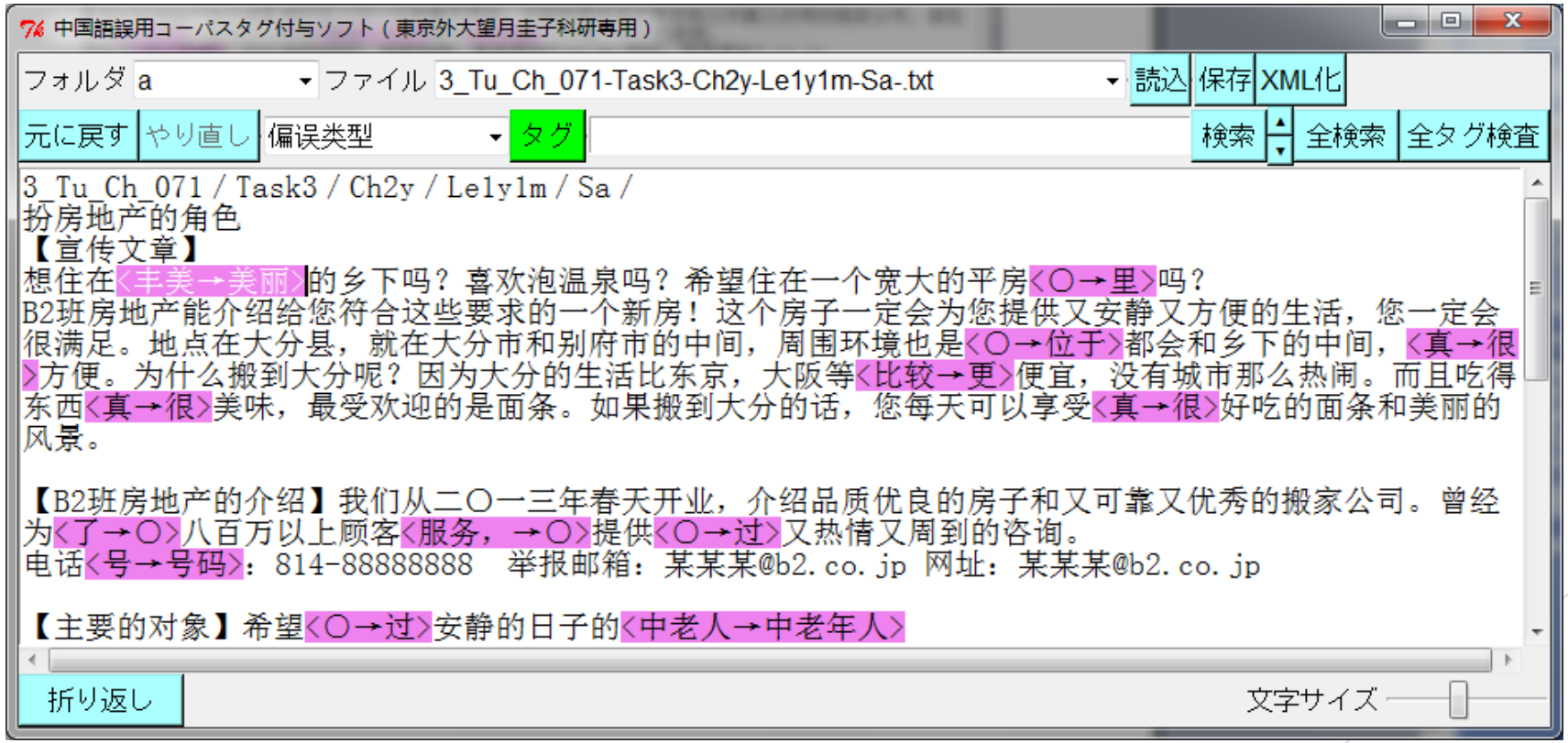


添加标签「误用类型」_1



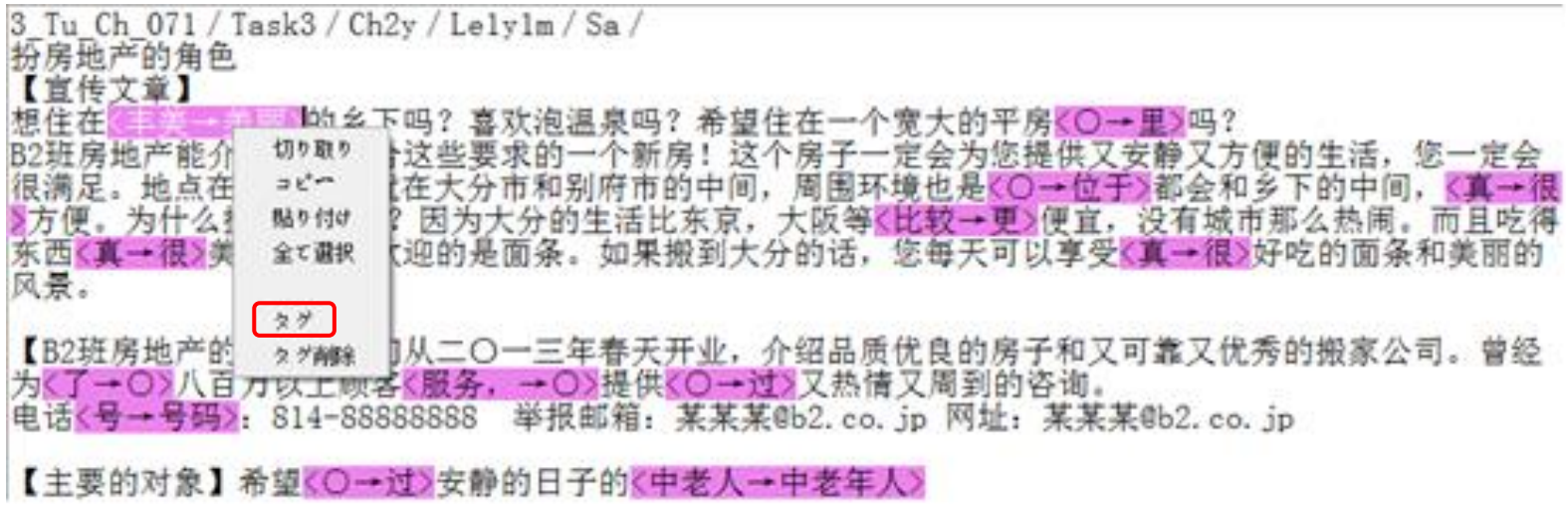
在标签列表中选择“偏误类型”。

再选中需要添加标签的部分（〈丰美→美丽〉）。



添加标签「误用类型」_2

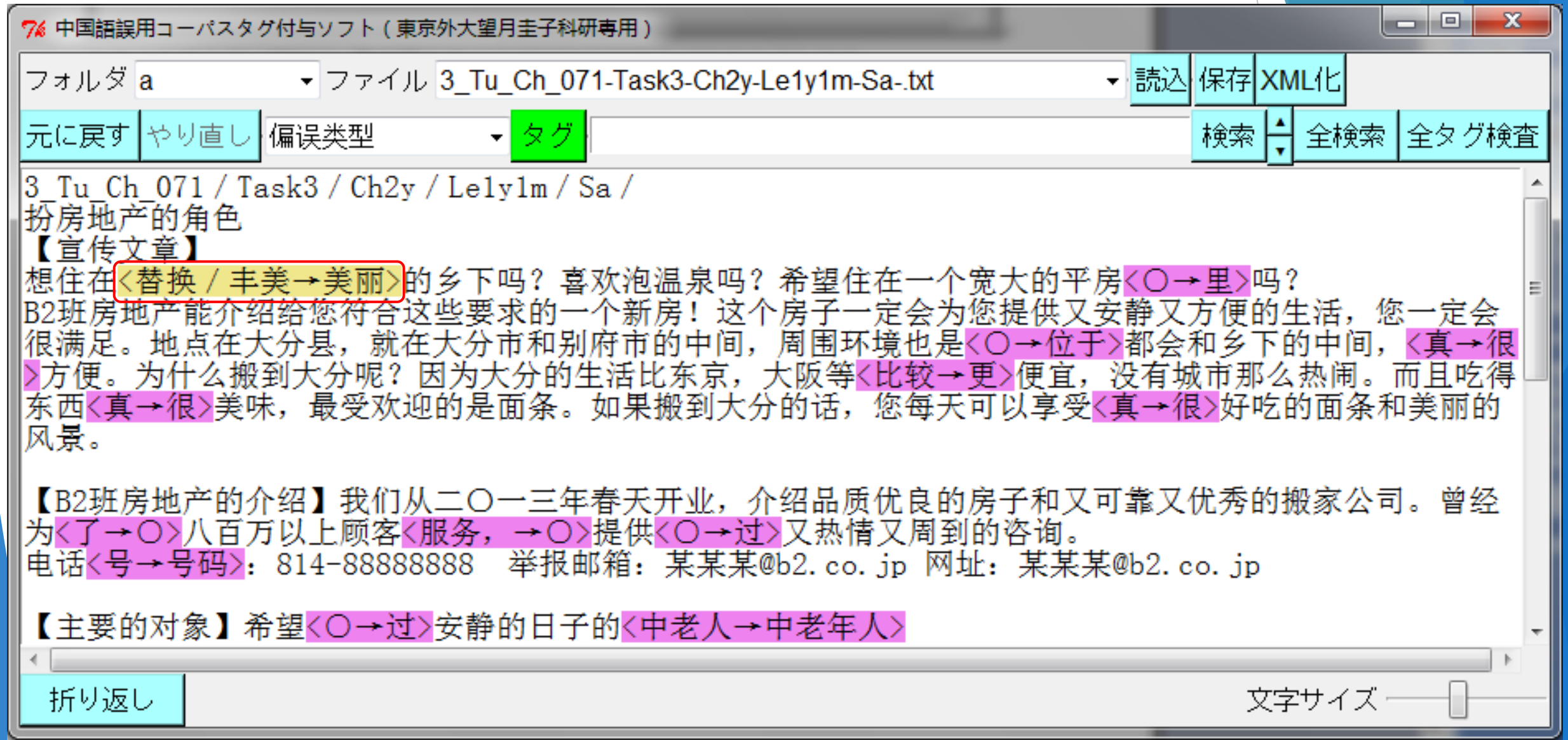
右键，然后选择「タグ」。



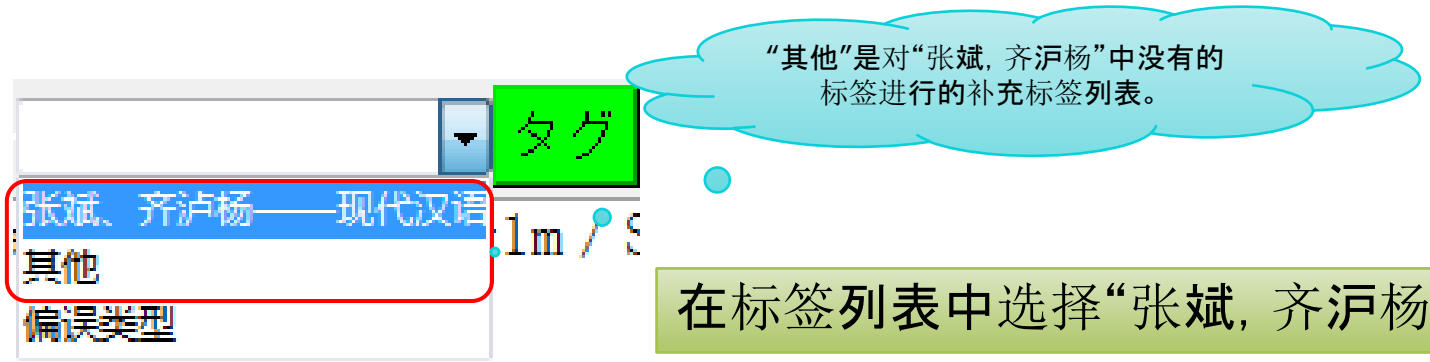
之后标签列表便会显示出来，选择「替换」，最后点击「タグ付与」即可。

添加标签「误用类型」_3

最后添加完的标签界面如下：

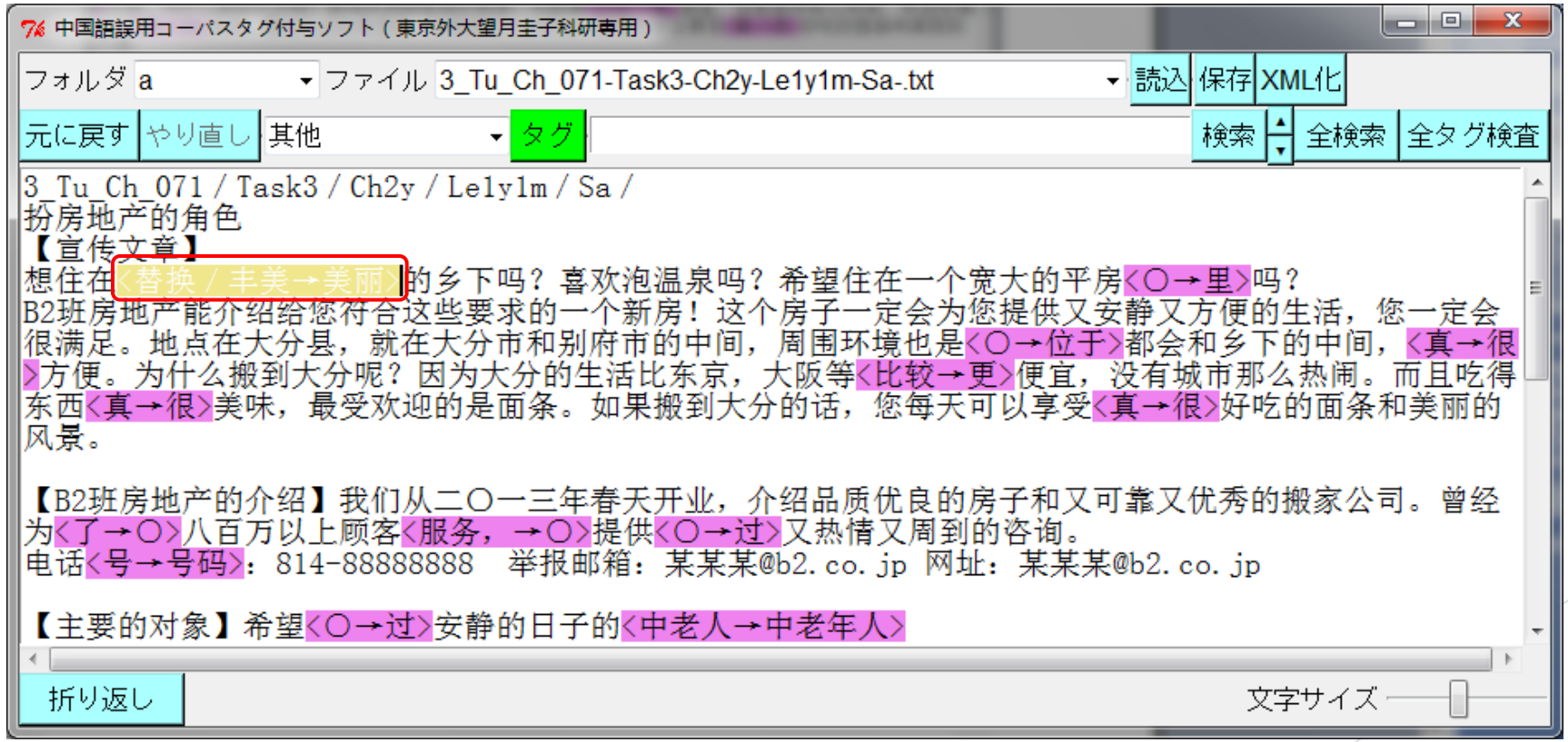


添加标签「语法标签」_1



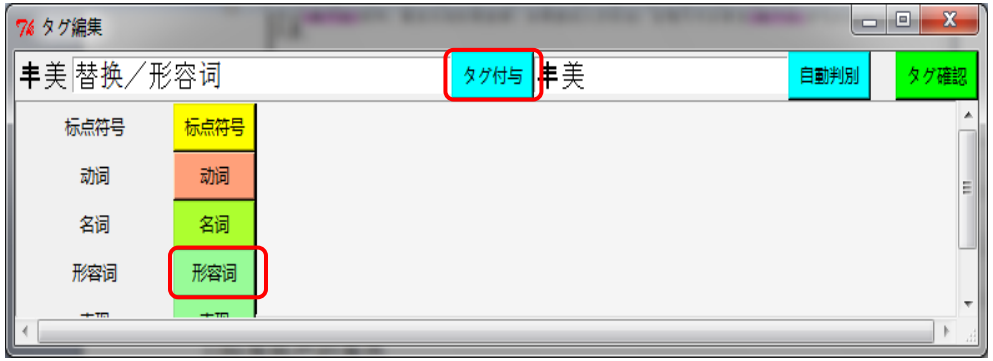
“其他”是对“张斌，齐沪杨”中没有的标签进行的补充标签列表。

在标签列表中选择“张斌，齐沪杨”或者“其他”。



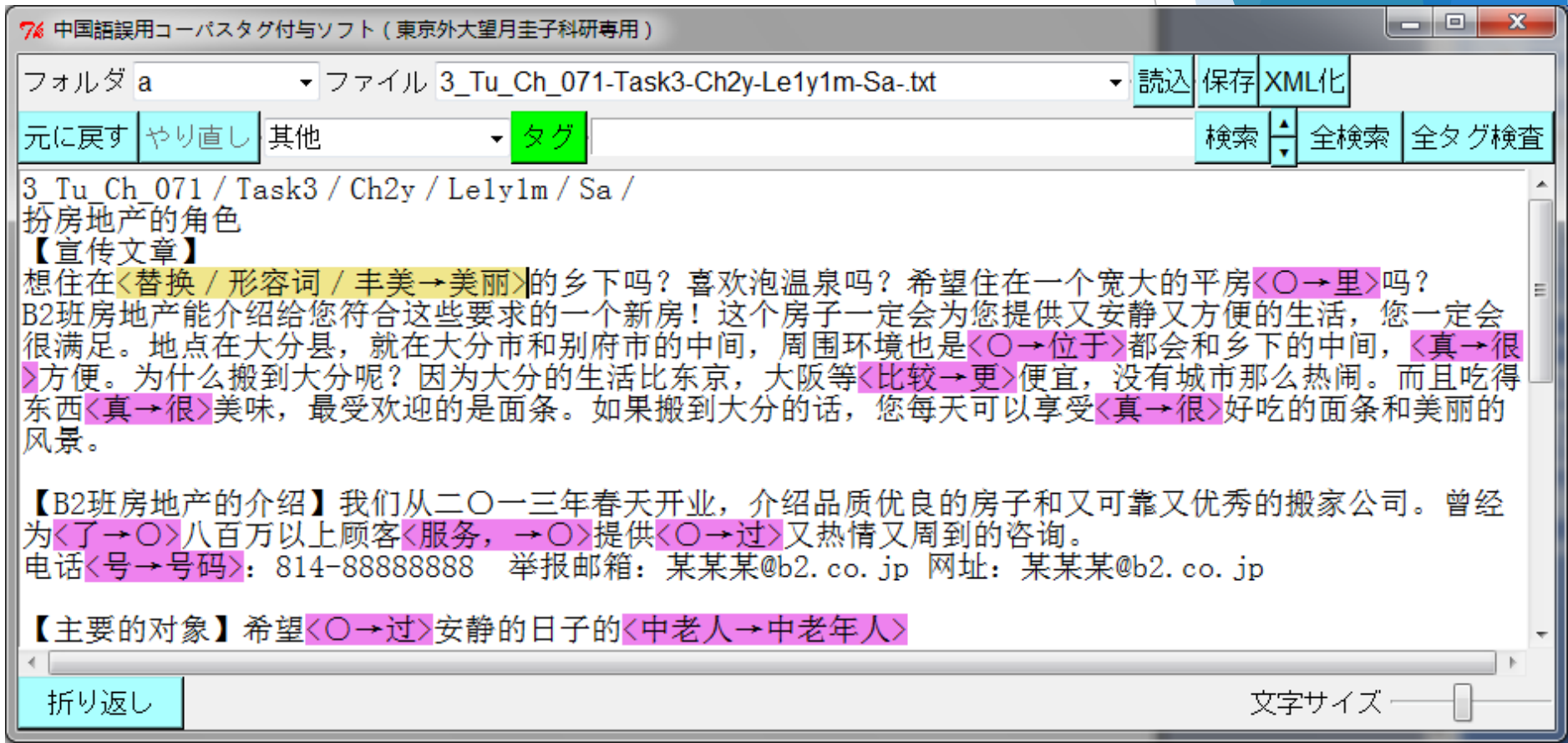
再次选择需要添加标签的部分（〈替换/丰美→美丽〉）。

添加标签「语法标签」_2

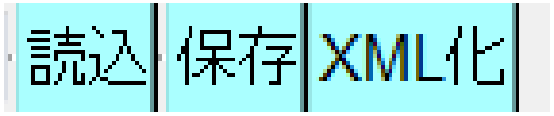


选择「形容词」后，点击「タグ付与」即可。

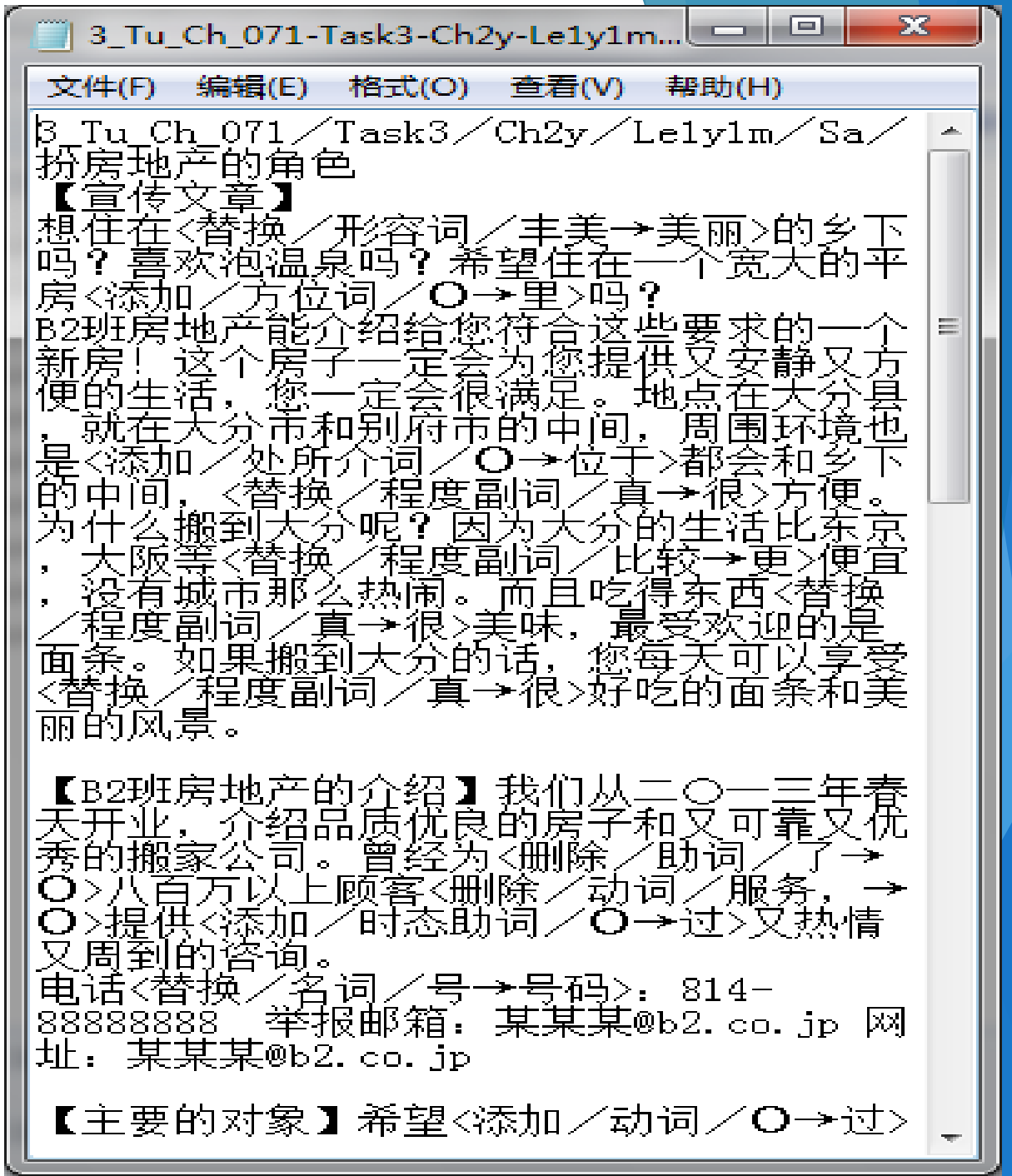
添加完两种标签的界面如图所示：



XML化



- 添加完所有标签后，点击「保存」即可完成文档的保存。
- 然后再点击「XML化」后则即刻生成XML网页文件。
- 最终将把生成的网页文件放入搜索引擎当中，即可完成误用语料库的构建工序。



谢谢！

2014. 6. 12