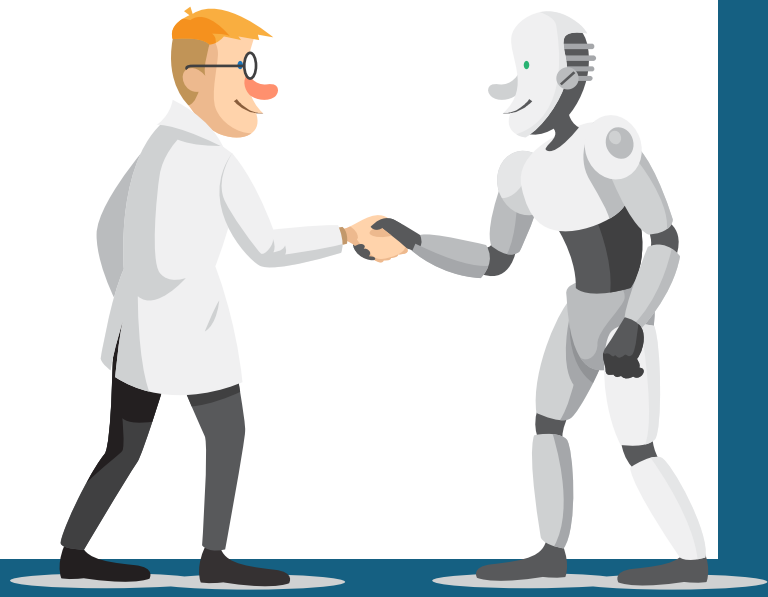
Week 5：语料库工具及语料库创建入门

黄婕

2024年10月16日

Class 3



本节内容

- 语料库基础
- 语料库的创建
- 术语的导出和对齐
- 本地单语、双语语料库工具
- 英文词性标注工具
- 语料库对齐工具
- 翻译记忆库的创建和导出

任务场景

为了完成一篇关于人工智能的科技文章的翻译，你需要搜索在线资料，然后使用语料库工具来提高翻译效率。具体包括：

1) 建立一个双语的语料库

2) 筛选出其中的术语，生成术语表

3) 基于双语的语料，生成双语翻译记忆库

请思考：你需要进行什么操作？过程中用到哪些工具？

本课学习的软件

语料库工具

Sketch Engine

强大的云端语料库工具

AntConc 4.1.1

北外ParaConc (tools)

过程工具

中文分词工具

英文词性标注Tree Tagger

Abbyy Aligner 2.0

1. 语料库基础

语料库是什么？ Corpus/Corpora

按照明确的目的和设计要求，根据语言学或翻译学原则，运用科学合理的技术方法，将一定规模的语言文本汇总而成的电子文本库。（管新潮，陶友兰，《语料库与翻译》，2017）

运用计算机技术，按照一定的语言学规则，根据特定的语言研究目的而大规模收集并存储在计算机中的真实语料，这些语料经过一定程度的标注，便于检索，可用于描述研究和实证研究。（王克非，《语料库翻译学探索》，2012）

语料库基础：分类


通用语料库/
专用语料库
文字语料库/
声音语料库
单语语料库/
平行语料库/
对比语料库
用于翻译研究的语料库/
用于翻译实践的语料库

常见的语料库

语言学或翻译学研究类语料库

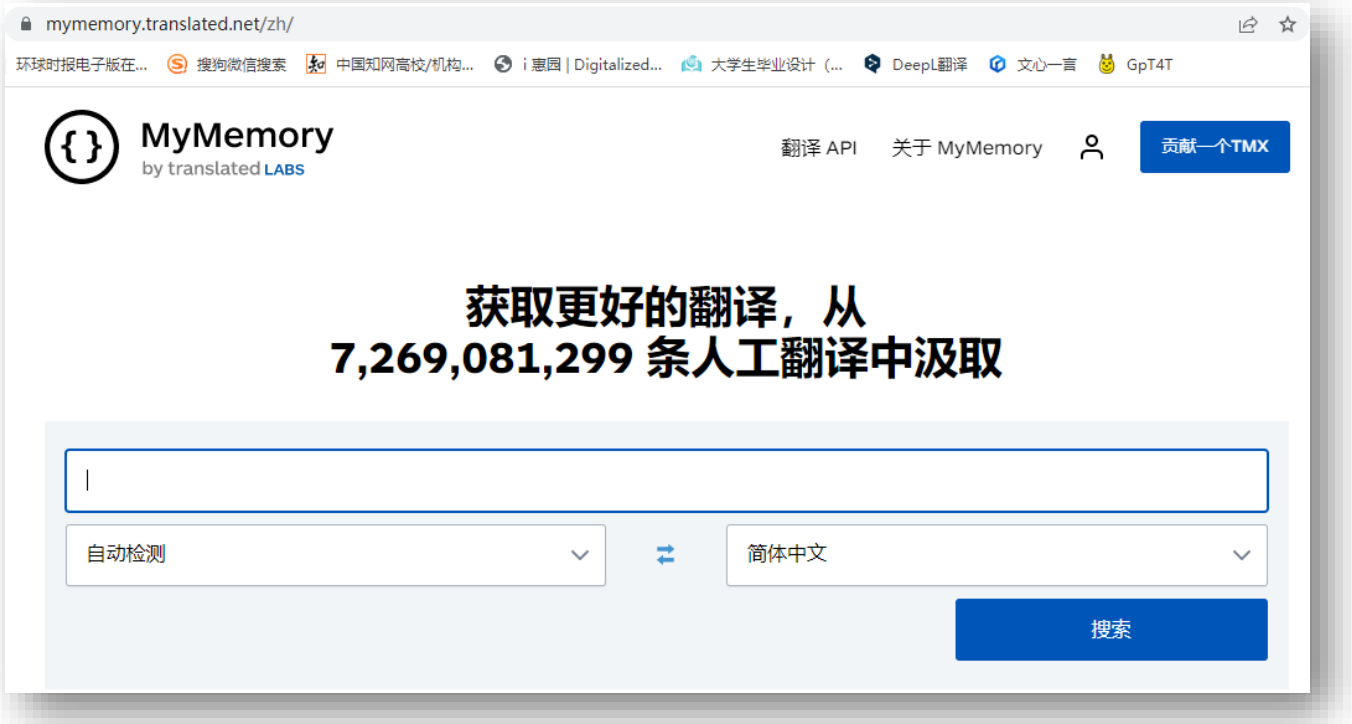
- 布朗（Brown）美国英语语料库（F. Nelson和H. Kucera，世界最早的计算机语料库）
- 上海交通大学科技英语语料库（JDEST，上海交通大学杨惠中）
- 翻译英语语料库（世界第一个翻译语料库，英国曼彻斯特大学，Mona Baker）
- 英国国家语料库（BNC）
- 美国当代英语语料库（COCA，美国杨百翰大学Mark Davies）
- 现代汉语语料库（中国国家语言文字委员会）
- 通用汉英对应语料库（北京外国语大学，王克非）
- 两岸三地英汉科普历时平行语料库（上海交通大学）
- 英汉医学平行语料库（上海交通大学，管新潮）
- 中国法律法规汉英平行语料库（绍兴文理学院）
- 汉学文史著作英汉平行语料库（山东师范大学，徐彬）

常见的语料库

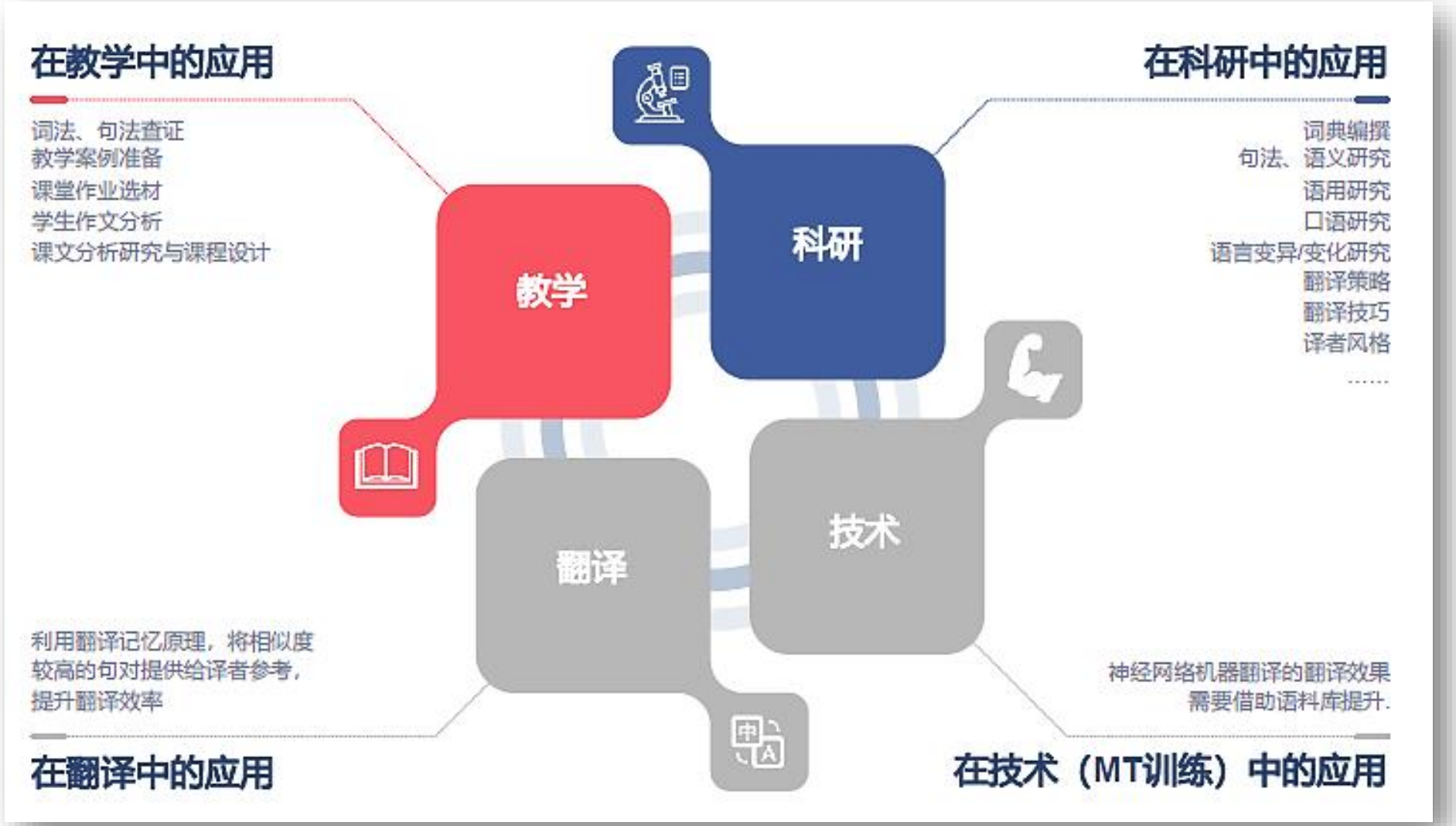
翻译实践应用类语料库

- 欧洲议会平行语料库（12亿单词，21种欧洲语言）
- TAUS Data (700亿单词，2200个语言对）
- MyMemory（https://mymemory.translated.net/）

TAUS
the language data network



语料库的功能


在教学中的应用
词法、句法查证
教学案例准备
课堂作业选材
学生作文分析
课文分析研究与课程设计
在科研中的应用
词典编撰
句法、语义研究
语用研究
口语研究
语言变异/变化研究
翻译策略
翻译技巧
译者风格
.....
教学
科研
翻译
技术
利用翻译记忆原理，将相似度较高的句对提供给译者参考，提升翻译效率
神经网络机器翻译的翻译效果需要借助语料库提升.
在翻译中的应用
在技术（MT训练）中的应用

语料库术语和功能

KWIC -- keyword in context

KWIC/concordance search 关键词索引： 特定词/短语的上下文使用案例

Word Frequency lists 词频列表：词或词组频繁出现的次数

Collocation analysis 搭配分析：某个词的前后搭配关系

WordList 词表：按照字母或词频顺序列出给定文本的词表，统计其文本特征。可以用于翻译准备阶段的术语提取

POS tagging 词性标注：(Part-of-speech tagging) 对语料库内的单词按照语义和上下文进行标记，即标注其词性

Keyword analysis 关键词分析

Term extraction 术语提取

双语平行语料库的应用

语料是基础

翻译风格研究

句法结构研究

跨文化传播研究、法律/商务话语研究、文化交际和形象研究、翻译研究以及翻译实践

构建语料检索、匹配、加工、利用、交易、开放的系统平台

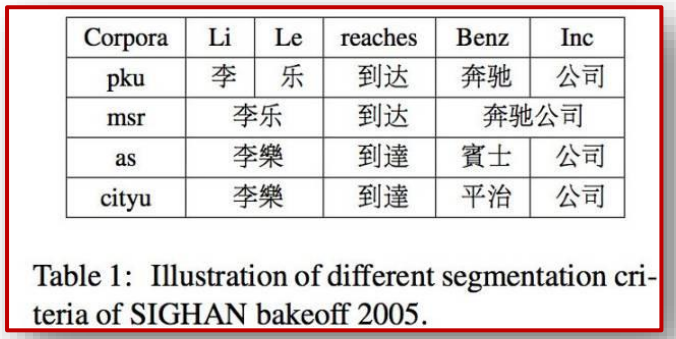
制定专业词表、术语（单语、双语）

统计字频、词频，编写教材

辅助写作，辅助翻译

训练机器翻译

语料库建设流程



Corpora	Li	Le	reaches	Benz	Inc
pku	李	乐	到达	奔驰	公司
msr	李乐		到达	奔驰公司	
as	李樂		到達	賓士	公司
cityu	李樂		到達	平治	公司

Table 1: Illustration of different segmentation criteria of SIGHAN bakeoff 2005.

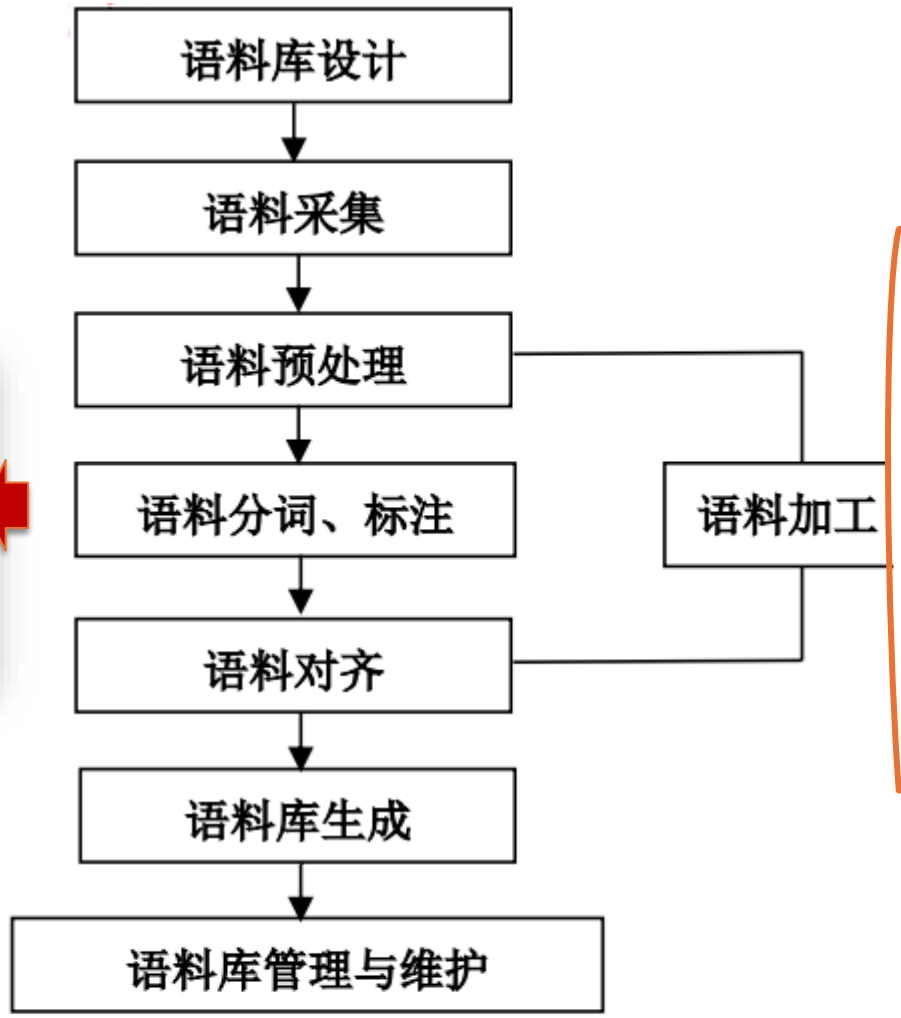


图 1：语料库建设流程图

什么是“对齐”？

- 在源语文本和目的语文本具体单位之间建立的对应关系。
- 可分为**词汇、语块、语句**、段落和篇章等层次对齐。

中国翻译协会，语料库通用技术规范，2018

2. 语料库的创建

任务场景

为了完成一篇关于人工智能的科技文章的翻译，你需要搜索在线资料，然后使用语料库工具来提高翻译效率。具体包括：

1) 建立一个双语的语料库

2) 筛选出其中的术语，生成一份术语对齐表。

先获取语料：

1. 客户提供参考文件
2. 网络搜索可靠资料

Step 1: 获取双语专业文档——搜索WIPO数据库

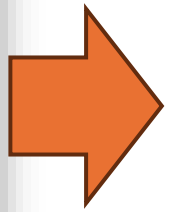
https://patentscope.wipo.int/

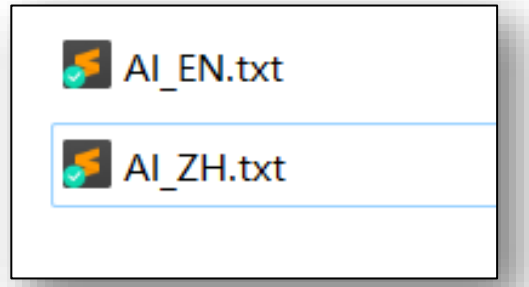
Step 1: 获取双语专业文档——保存为本地文档

摘要

[EN] The embodiment of the invention provides a training method of a large language model and a text processing method based on the large language model, and relates to the fields of artificial intelligence, cloud technology, natural language processing, machine learning and the like, in particular to a language model in a pre-training language model. The method comprises the following steps: acquiring a training set and a pre-training language model corresponding to each task in a plurality of natural language processing tasks in the same target field, and acquiring a second feature extraction network corresponding to each task; and repeatedly performing training operation on the second feature extraction network corresponding to the task based on the training set corresponding to the task until a training ending condition is met, obtaining a trained second feature extraction network corresponding to the task, and obtaining the target target based on the pre-training language model and the trained second feature extraction network corresponding to each task. And obtaining a target large language model of the target domain. Based on the method, the accuracy of the text processing result output by the large language model can be improved.

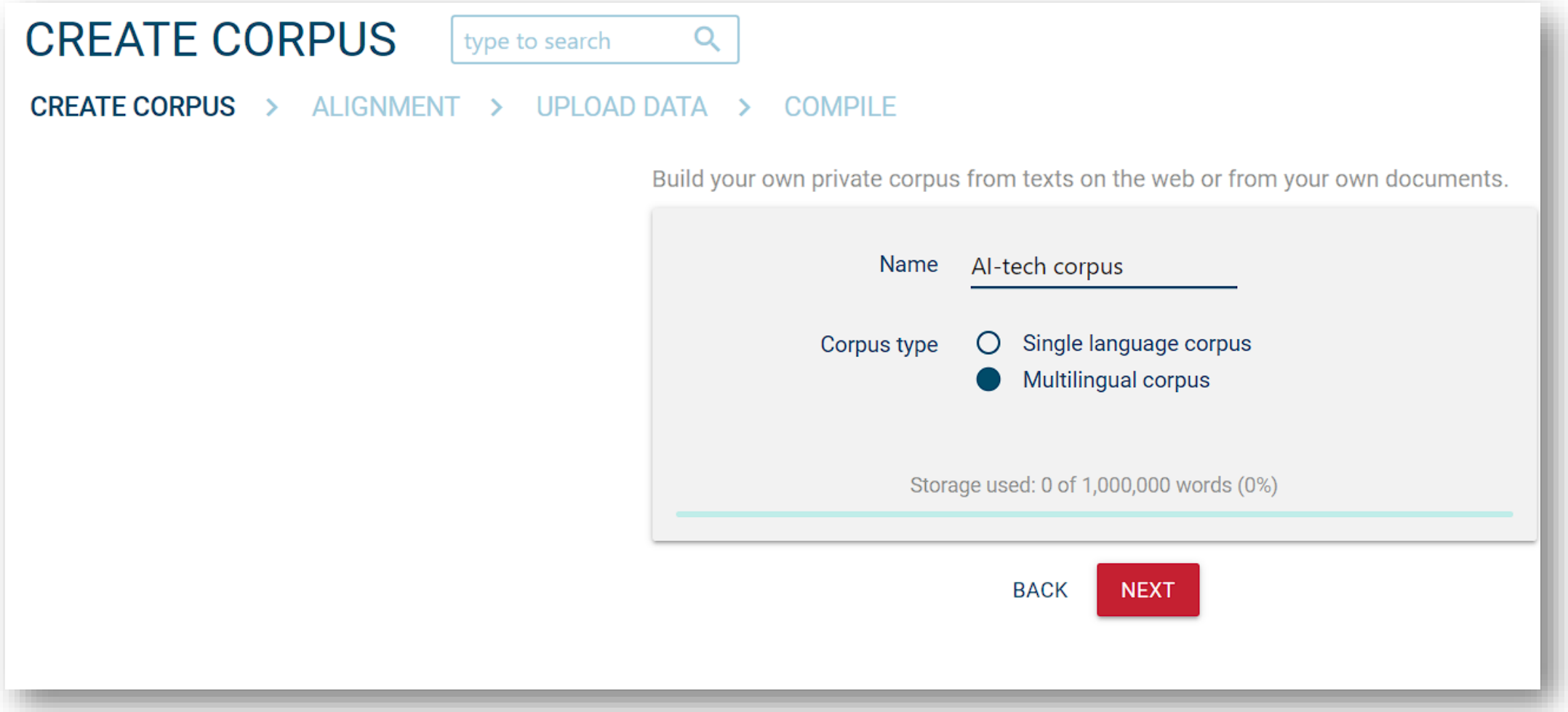
[ZH] 本申请实施例提供了一种大语言模型的训练方法及基于大语言模型的文本处理方法，涉及人工智能、云技术、自然语言处理及机器学习等领域，尤其涉及预训练语言模型中的语言模型。该方法包括：获取同一目标领域的多个自然语言处理任务中每一任务对应的训练集和预训练语言模型，获取每一任务对应的第二特征提取网络，对于每一任务，基于该任务对应的训练集对该任务对应的第二特征提取网络重复执行训练操作，直至满足训练结束条件，得到该任务对应的训练后的第二特征提取网络，基于所述预训练语言模型和各所述任务对应的训练后的第二特征提取网络，得到所述目标领域的目标大语言模型。基于该方法，可以提高大语言模型输出的文本处理结果的准确性。



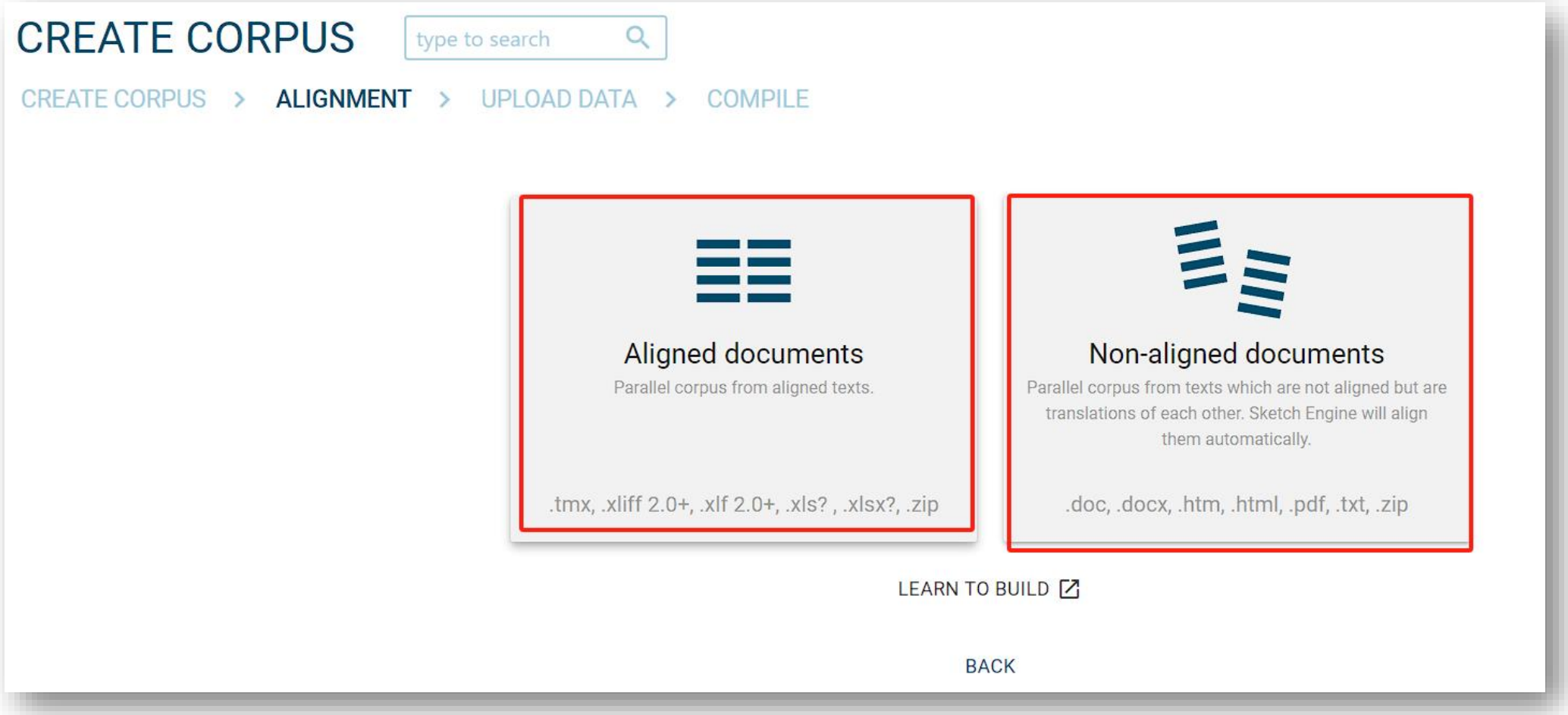


Step 2 创建语料库，导入双语文档

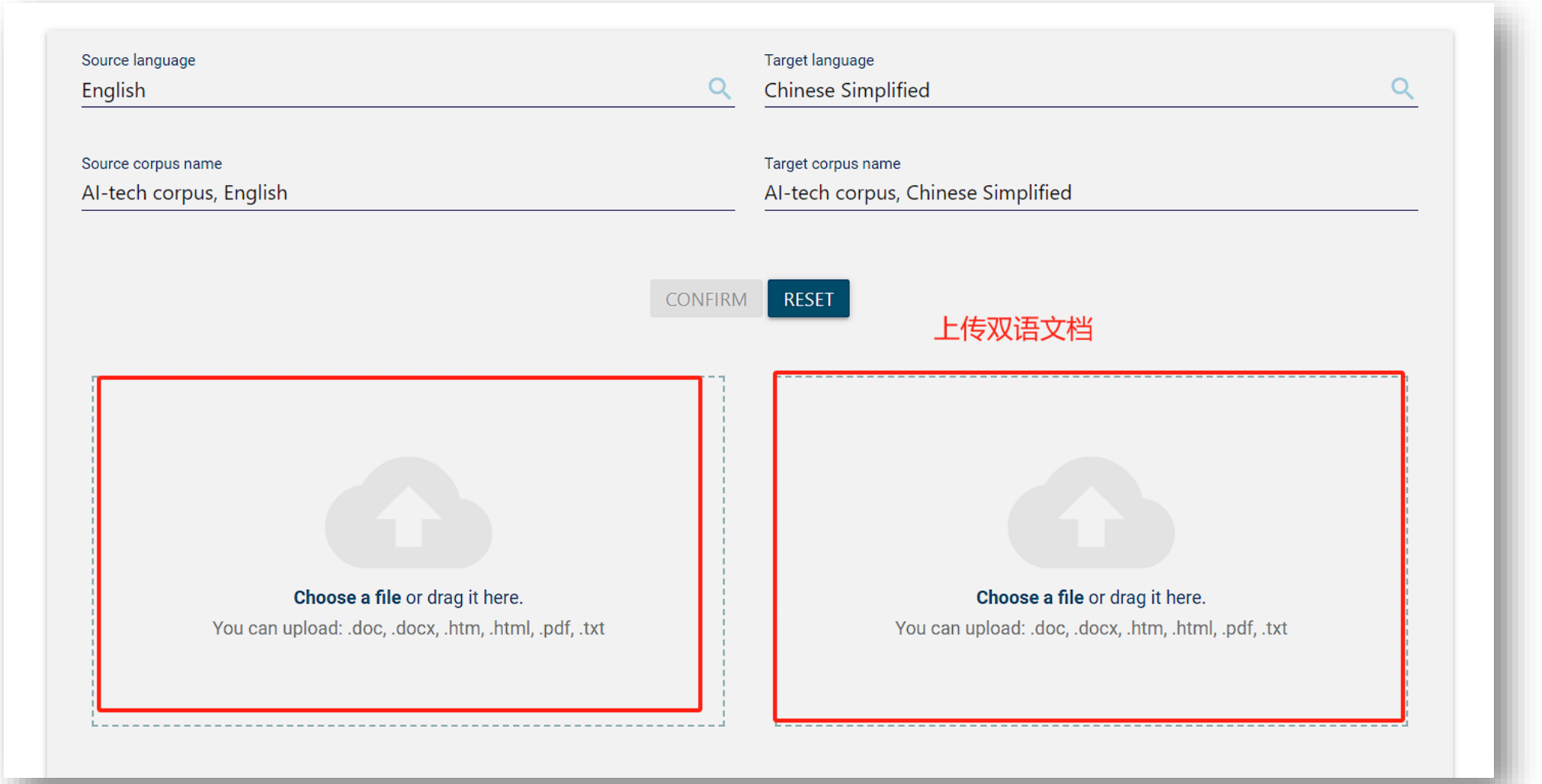
- https://auth.sketchengine.eu/#login



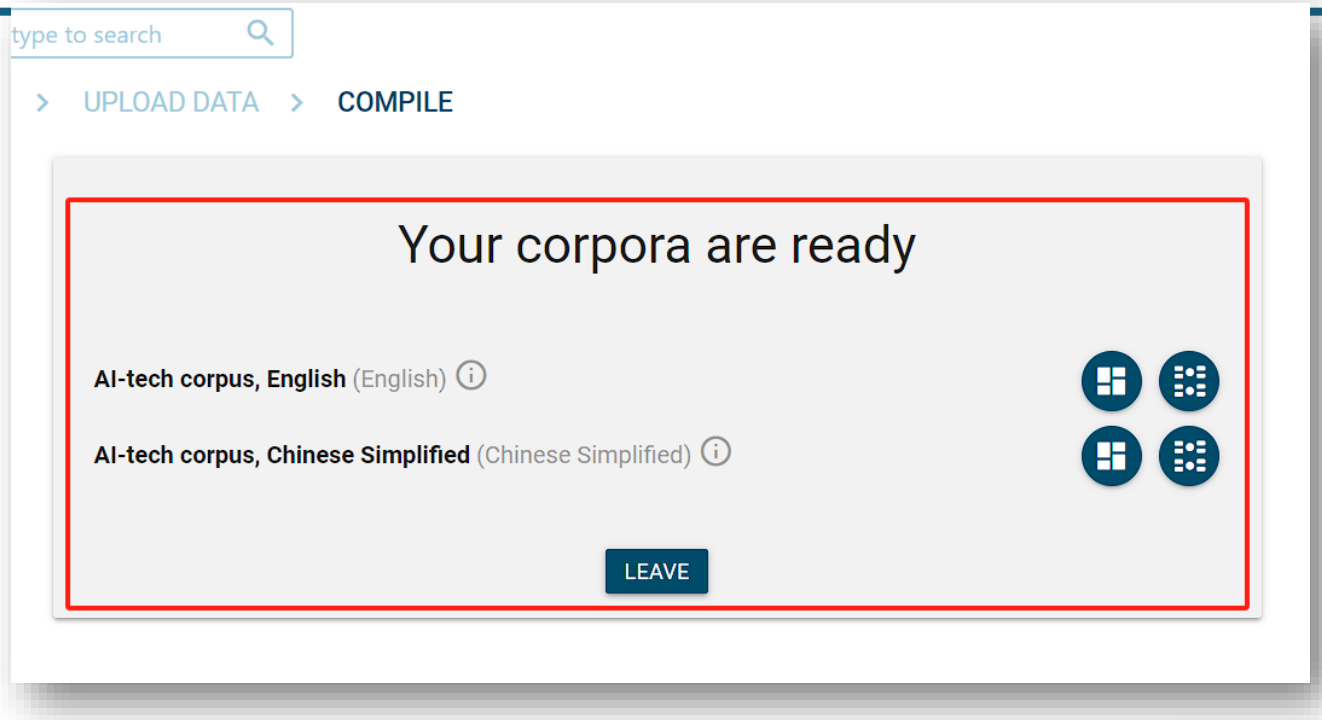
Step 2 创建语料库，导入双语文档



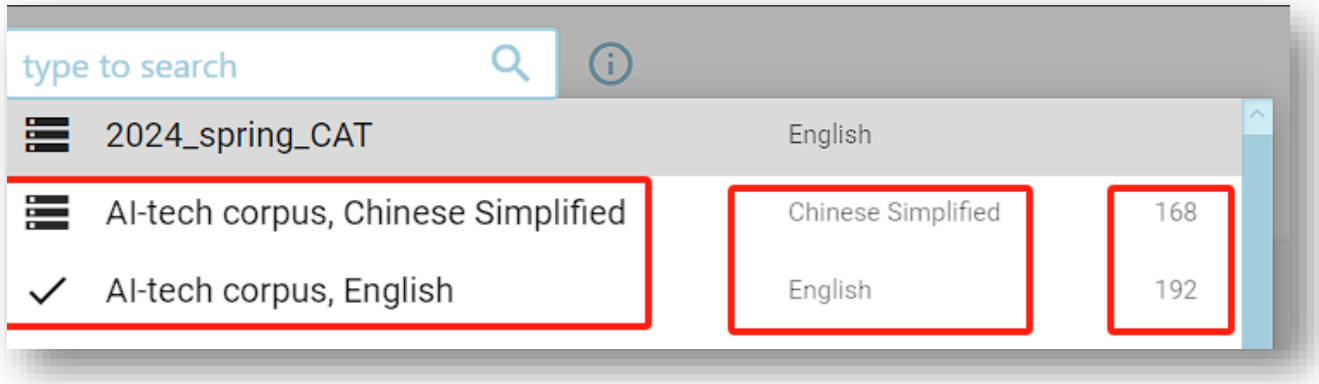
Step 2 创建语料库，导入双语文档



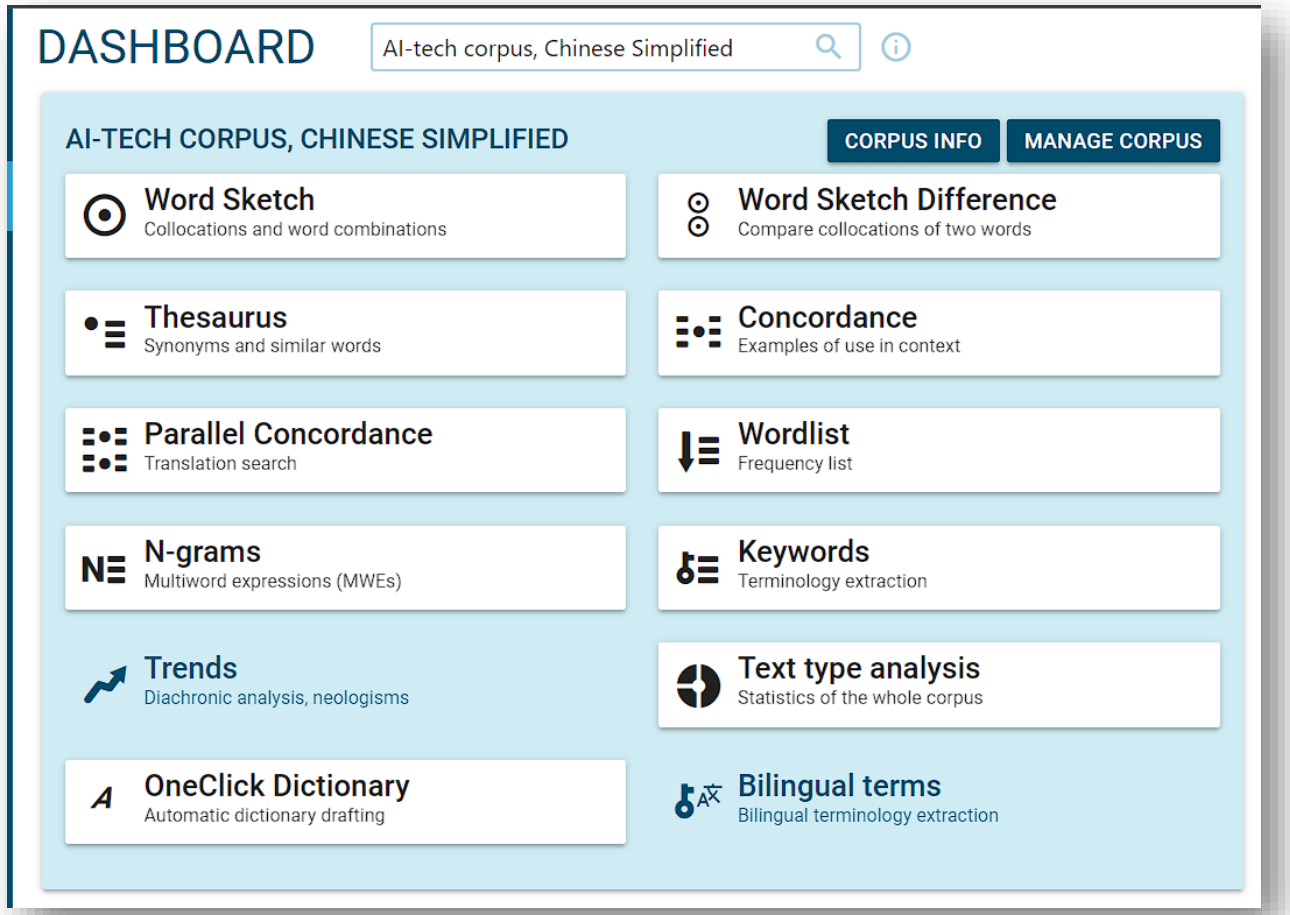
Step 2 创建语料库，导入双语文档



创建成功



Sketch engine 的基本功能



Word sketch 词汇素描:

展示一个单词的语法和搭配行为，包括常见的修饰词、宾语、主语等。

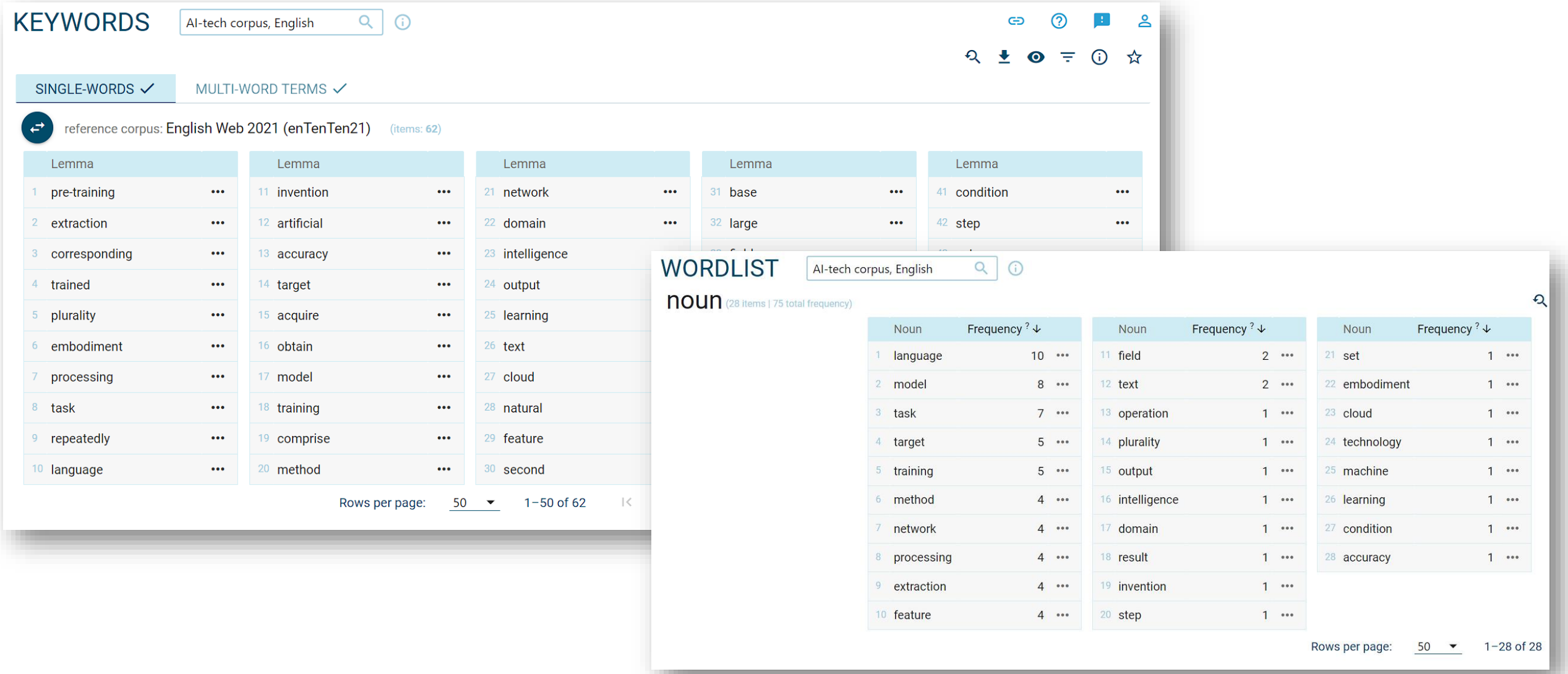
Thesaurus 同义词：

用于查找同义词，特别适用于写作时遇到词穷的情况。

Wordlist 词频列表

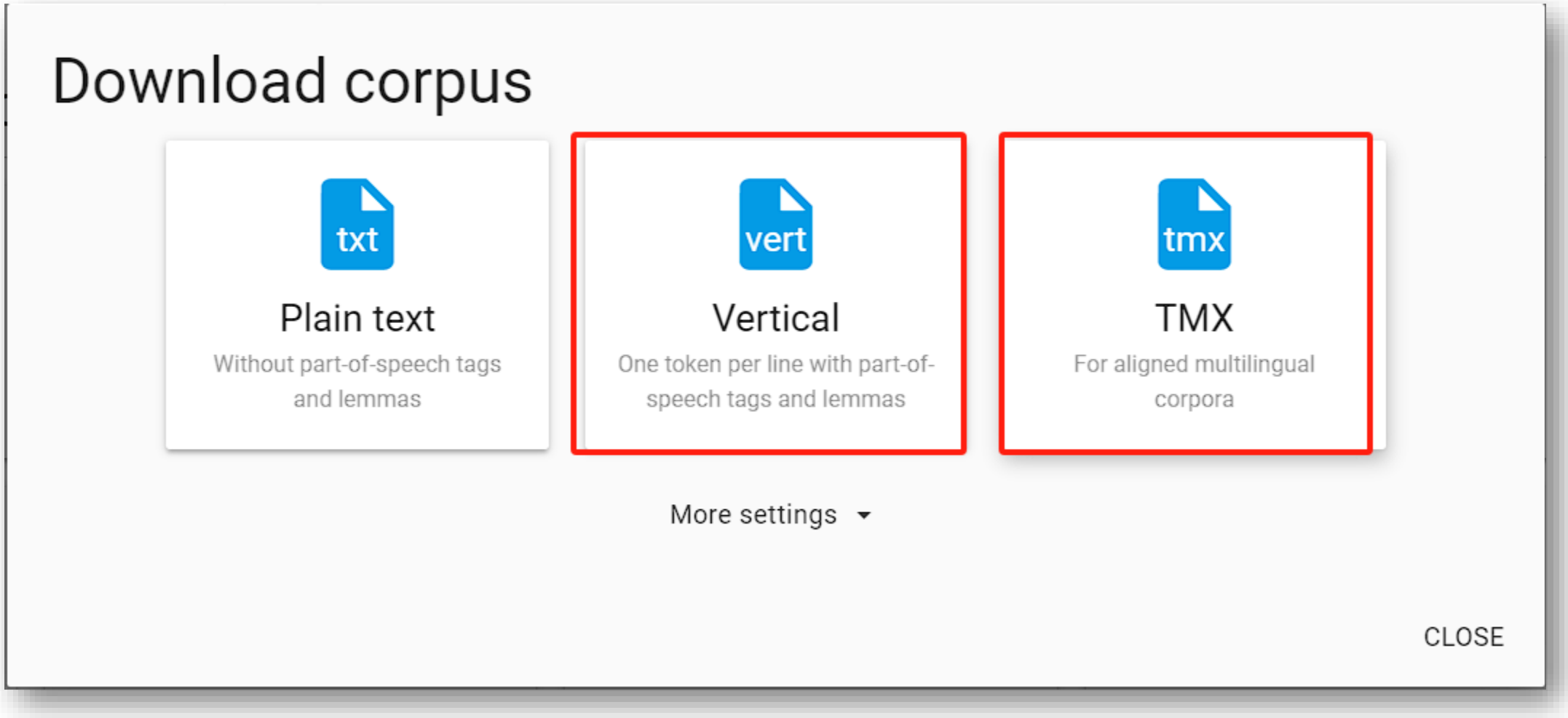
Keywords 关键词术语表

Sketch engine: 关键词、词频（ 自动词性标注)



3. 术语的导出和对齐

Step 3 导出术语、导出双语记忆库文件



Step 3 导出术语、导出双语记忆库文件

vert 文件：

术语的词性标注

ai_tech_corpus_english.vert

```
<doc id="file32377953" filename="AI_EN.txt" parent_folder="upload">
<s>
The DT the-x
embodiment NN embodiment-n
of IN of-i
the DT the-x
invention NN invention-n
provides VVZ provide-v
a DT a-x
training NN training-n
method NN method-n
of IN of-i
a DT a-x
large JJ large-j
language NN language-n
model NN model-n
and CC and-c
a DT a-x
text NN text-n
processing NN processing-n
method NN method-n
based VVN base-v
on IN on-i
the DT the-x
large JJ large-j
language NN language-n
model NN model-n
<g/>
```

术语的词性标注+详细信息

Step 3 导出术语、导出双语记忆库文件

tmx 文件：

双语翻译记忆库

ai_tech_corpus_english.tmx

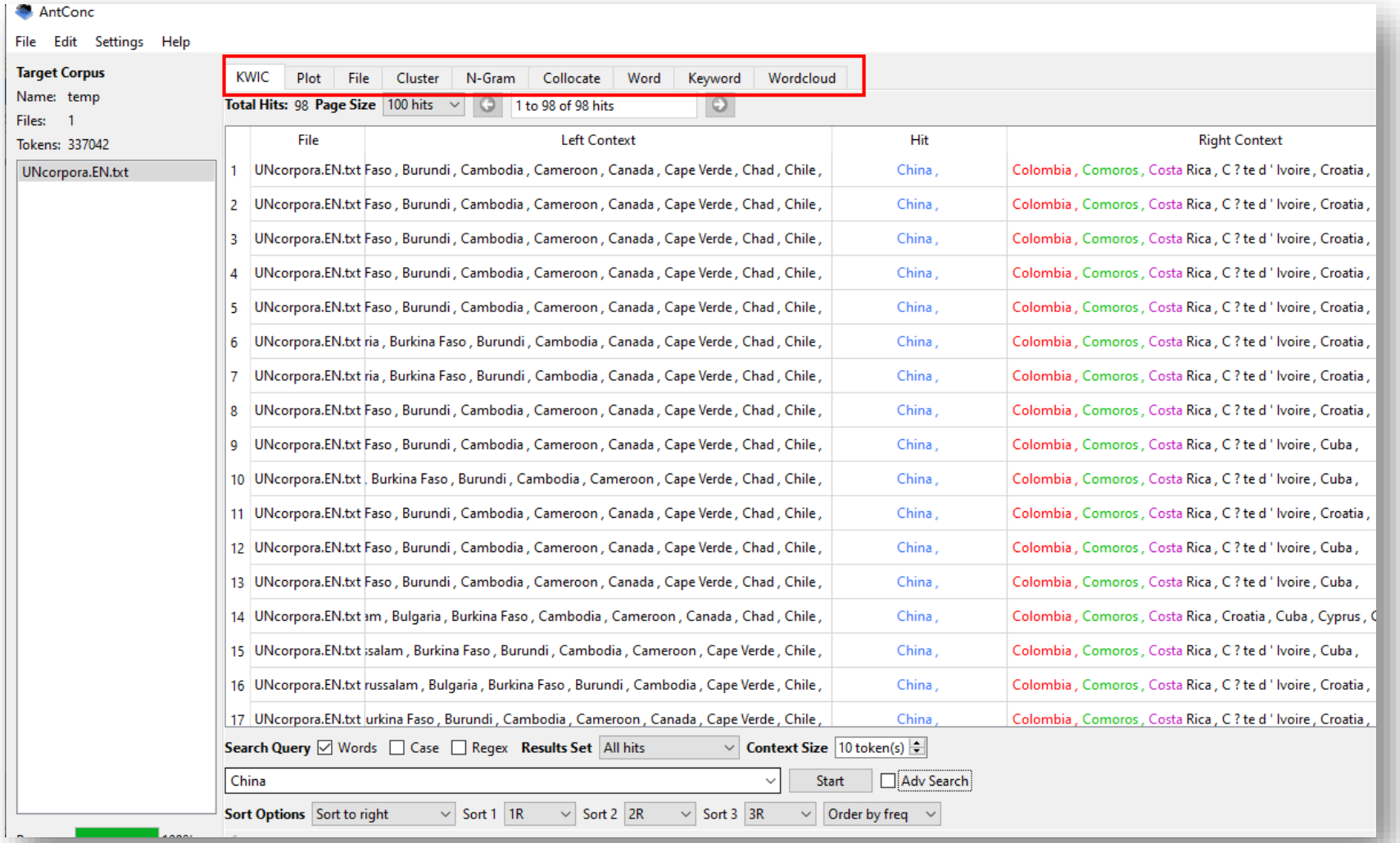
```
<?xml version="1.0" encoding="UTF-8" standalone="no" ?>
<tmx version="1.4"><header /><body>
<tu>
<tuv xml:lang="en"><seg>The embodiment of the invention provides a training method of a large language model and a text processing method based on the large language model, and relates to the fields of artificial intelligence, cloud technology, natural language processing, machine learning and the like, in particular to a language model in a pre-training language model.</seg></tuv>
<tuv xml:lang="zh-Hans"><seg>本 申请 实施 例 提供 了 一 种 大 语言 模型 的 训练 方法 及 基于 大 语言 模型 的 文本 处理 方法 ， 涉及 人工 智能 、 云 技术 、 自然 语言 处理 及 机器 学习 等 领域 ， 尤其 涉及 预 训练 语言 模型 中 的 语言 模型 。</seg></tuv>
</tu>
<tu>
<tuv xml:lang="en"><seg>The method comprises the following steps: acquiring a training set and a pre-training language model corresponding to each task in a plurality of natural language processing tasks in the same target field, and acquiring a second feature extraction network corresponding to each task; and repeatedly performing training operation on the second feature extraction network corresponding to the task based on the training set corresponding to the task until a training ending condition is met, obtaining a trained second feature extraction network corresponding to the task, and obtaining the target target based on the pre-training language model and the trained second feature extraction network corresponding to each task.</seg></tuv>
<tuv xml:lang="zh-Hans"><seg>该 方法 包括 ： 获取 同一 目标 领域 的 多 个 自然 语言 处理 任务 中 每 一 任务 对应 的 训练 集 和 预 训练 语言 模型 ， 获取 每 一 任务 对应 的 第二 特征 提取 网络 ， 对于 每 一 任务 ， 基于 该 任务 对应 的 训练 集 对 该 任务 对应 的 第二 特征 提取 网络 重复 执行 训练 操作 ， 直 至 满足 训练 结束 条件 ， 得到 该 任务 对应 的 训练 后 的 第二 特征 提取 网络 ， 基于 所述预 训练 语言 模型 和 各 所 述 任务 对应 的 训练 后 的 第二 特征 提取 网络 ， 得到 所述 目标 领域 的 目标 大 语言 模型 。</seg></tuv>
</tu>
```

4. 本地单语、双语语料库工具

单语语料库工具：AntConc

http://www.laurenceanthony.net/software/antconc/

UTF-8编码：
通用、国际化



单语语料库工具：WordSmith

优点：

1.功能丰富：词汇检索、词频统计、共现分析等

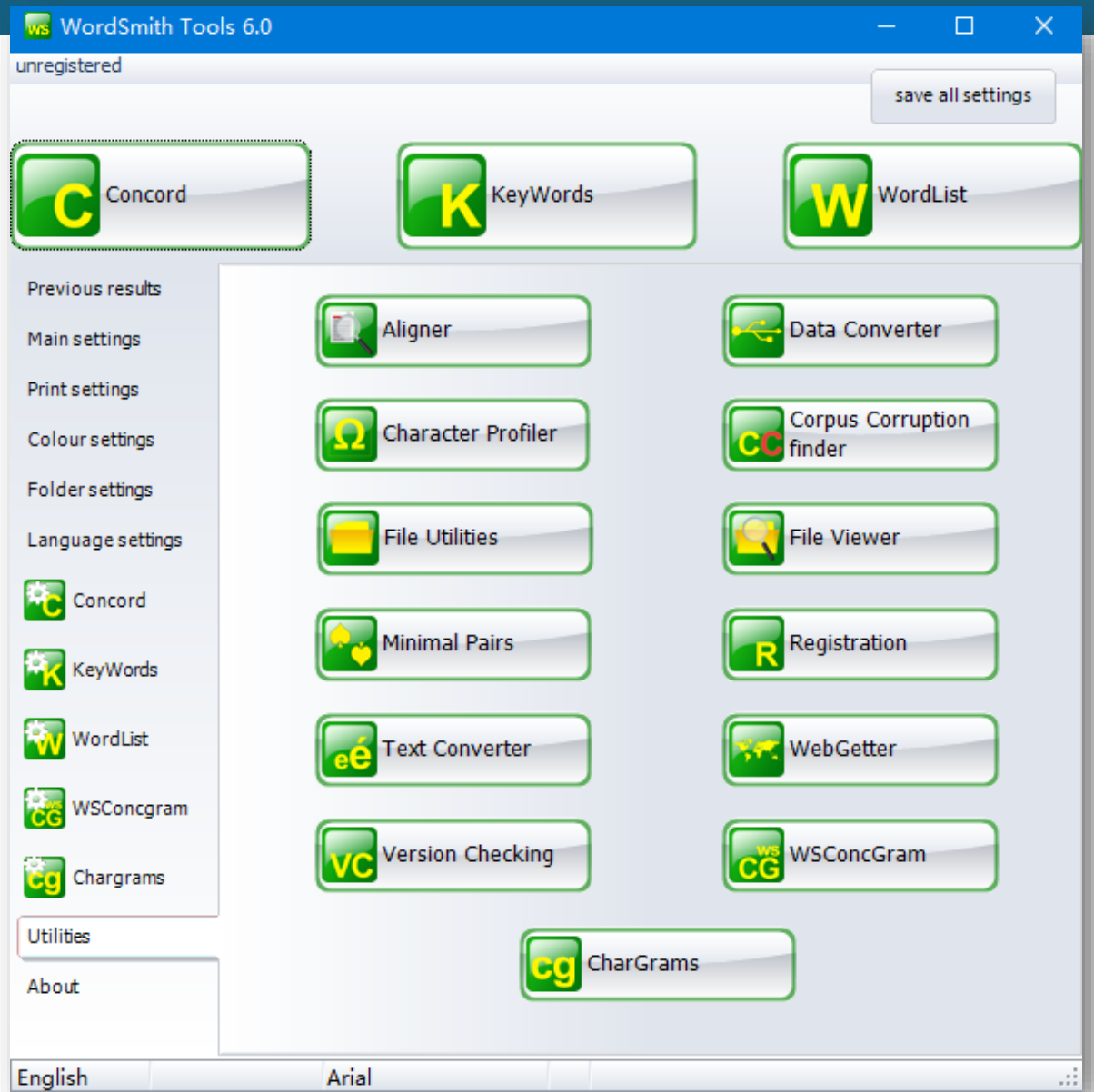
2.界面友好：软件界面布局合理，操作相对简便

3.支持大文件处理：可以处理大型语料库。

缺点：

1.商业软件：需要付费使用

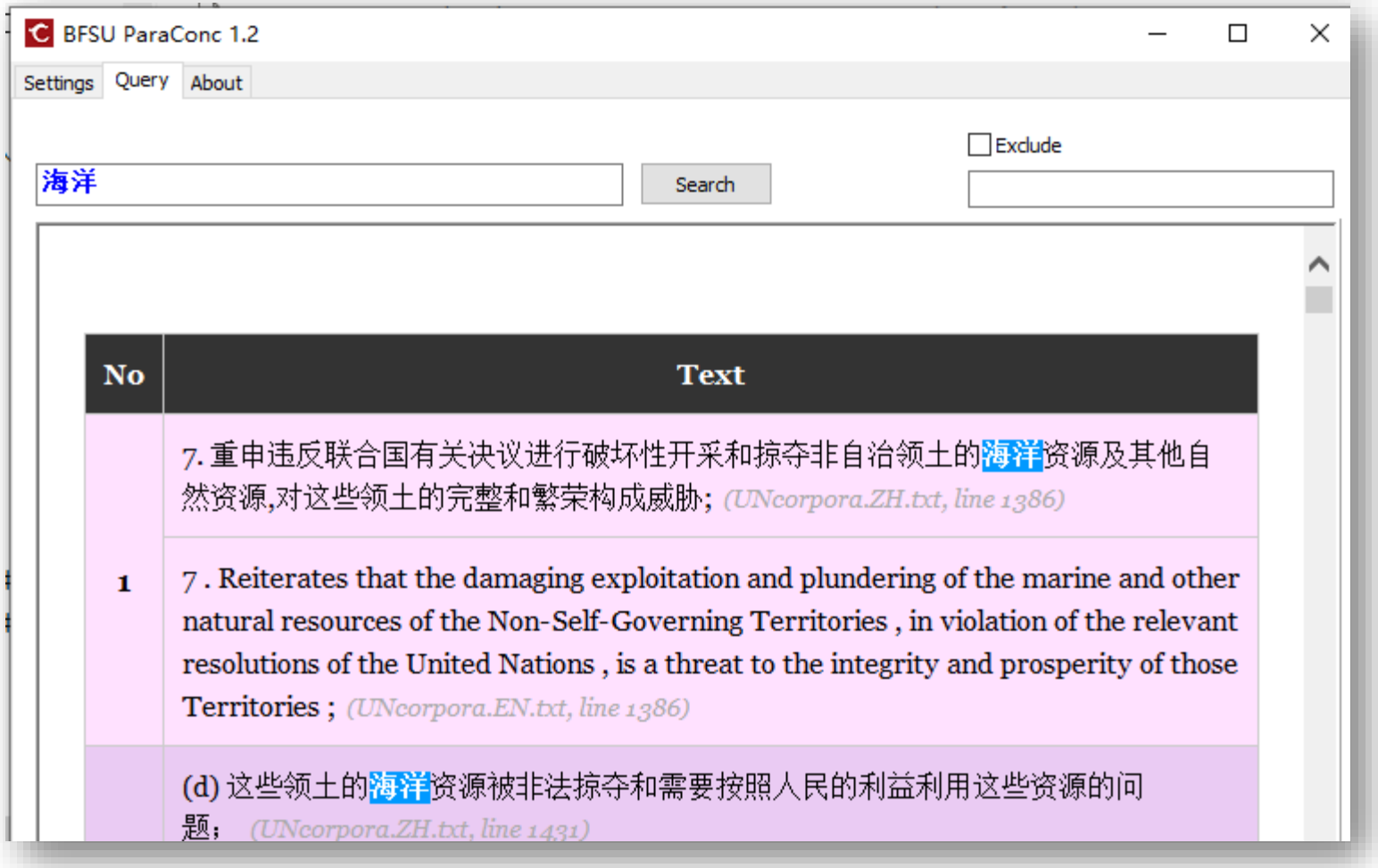
2.功能细节不尽完善：虽然功能丰富，但细节方面可能不如其他免费的AntConc。



双语语料库工具： ParaConc

ANSI编码：
需转换、有局限性

但是可以创建双语语料库



5. 中文分词、英文词性标注工具

中文分词的目的

将自然语言分解为最小的单元（词语），为以下活动做准备：

- 生成词典
- 多语语料对齐

常见中文分词的工具（Python包)

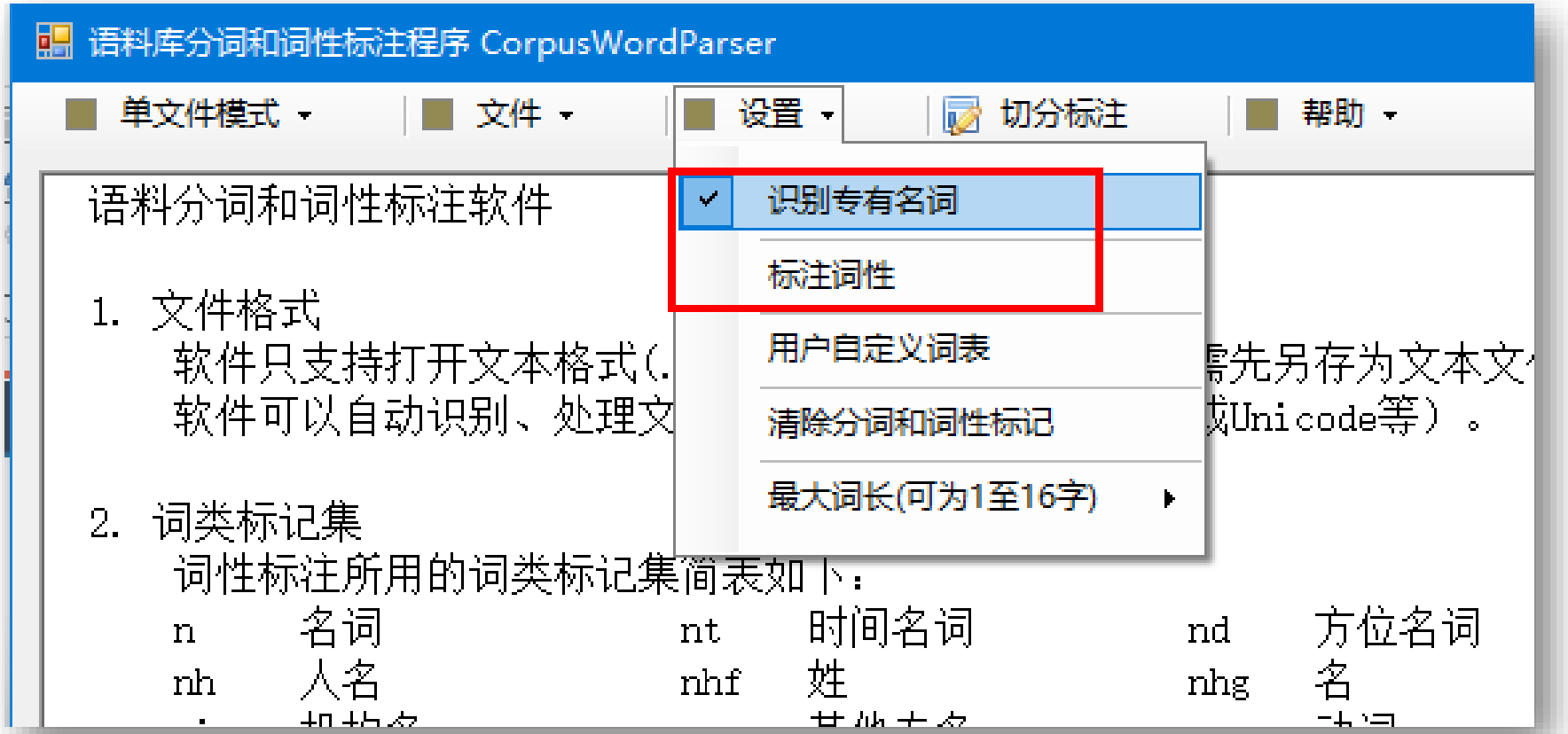
Jieba：中文分词+词性标注（热门工具）

SnowNLP：中文分词+词性标注、情感分析、文本分类等

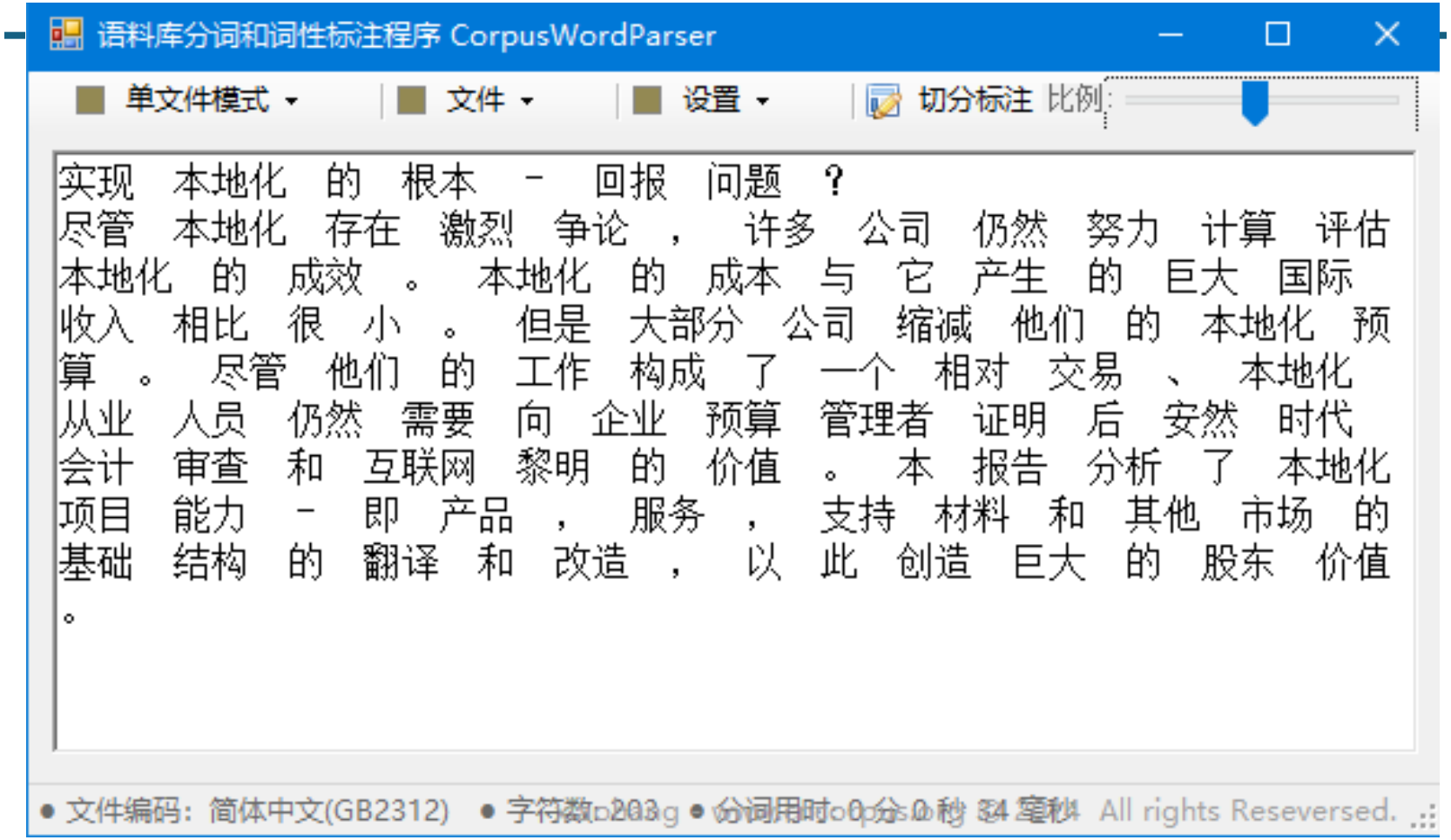
PkuSeg：多领域分词，个性化的预训练模型。

THULAC：中文分词+词性标注。清华大学自然语言处理与社会人文计算实验室。

中文自动分词工具：CorpusWordParser.exe



中文自动分词工具：CorpusWordParser.exe



中文语料库在线分词与标注工具，分词后的文件为ANSI编码的TXT文件

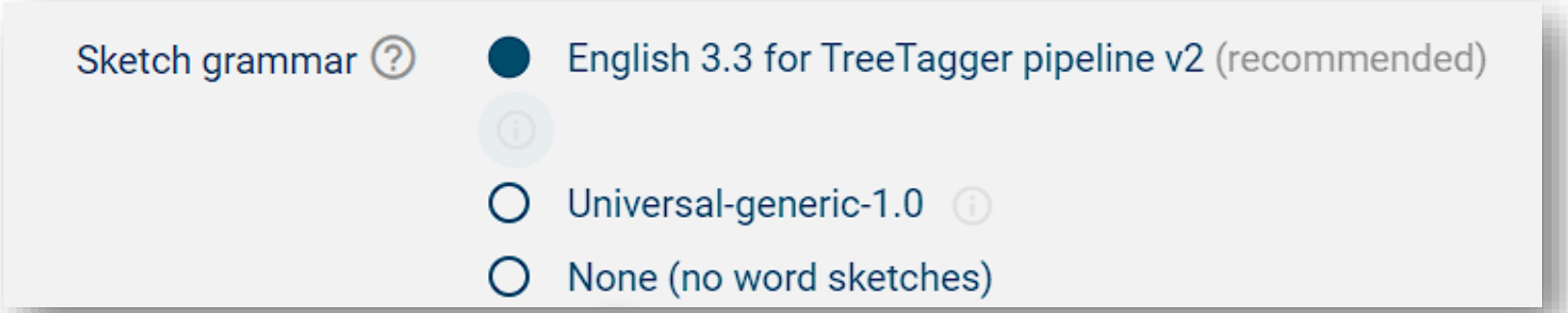
英文词性标注工具：Tree Tagger


```
Getting_VVG to_TO the_DT Bottom_NP of_IN Localization_NP -_NN
What_WP Is_VBZ the_DT Payback_NN ?_SENT
Despite_IN compelling_JJ arguments_NNS for_IN localization_NN ,_,
many_JJ firms_NNS are_VBP still_RB struggling_VVG with_IN
figuring_VVG out_RP how_WRB to_TO justify_VV the_DT effort_NN
._SENT
The_DT cost_NN of_IN localization_NN happens_VVZ to_TO be_VB
very_RB small_JJ compared_VVN to_TO the_DT big_JJ
international_JJ revenue_NN it_PP can_MD help_VV generate_VV
._SENT
But_CC most_JJS firms_NNS shortchange_VV their_PP$
localization_NN budgets_NNS ._SENT
Even_RB though_IN their_PP$ work_NN constitutes_VVZ a_DT
relative_JJ bargain_NN ,_, localization_NN practitioners_NNS
still_RB have_VHP to_TO prove_VV their_PP$ value_NN to_TO
corporate_JJ budgeters_NNS mindful_JJ of_IN post-Enron_NN
accounting_NN scrutiny_NN and_CC the_DT morning-after_JJ
internet_NN malaise_NN ._SENT
```

ANSI编码的TXT文件

英文词性标注工具：Tree Tagger

其他调用 tree tagger 的方式



The following steps are necessary to install the TreeTagger (see below for the Windows version). Download the files by right-clicking on the link. Then select "save file as". All files should be stored in the same directory.

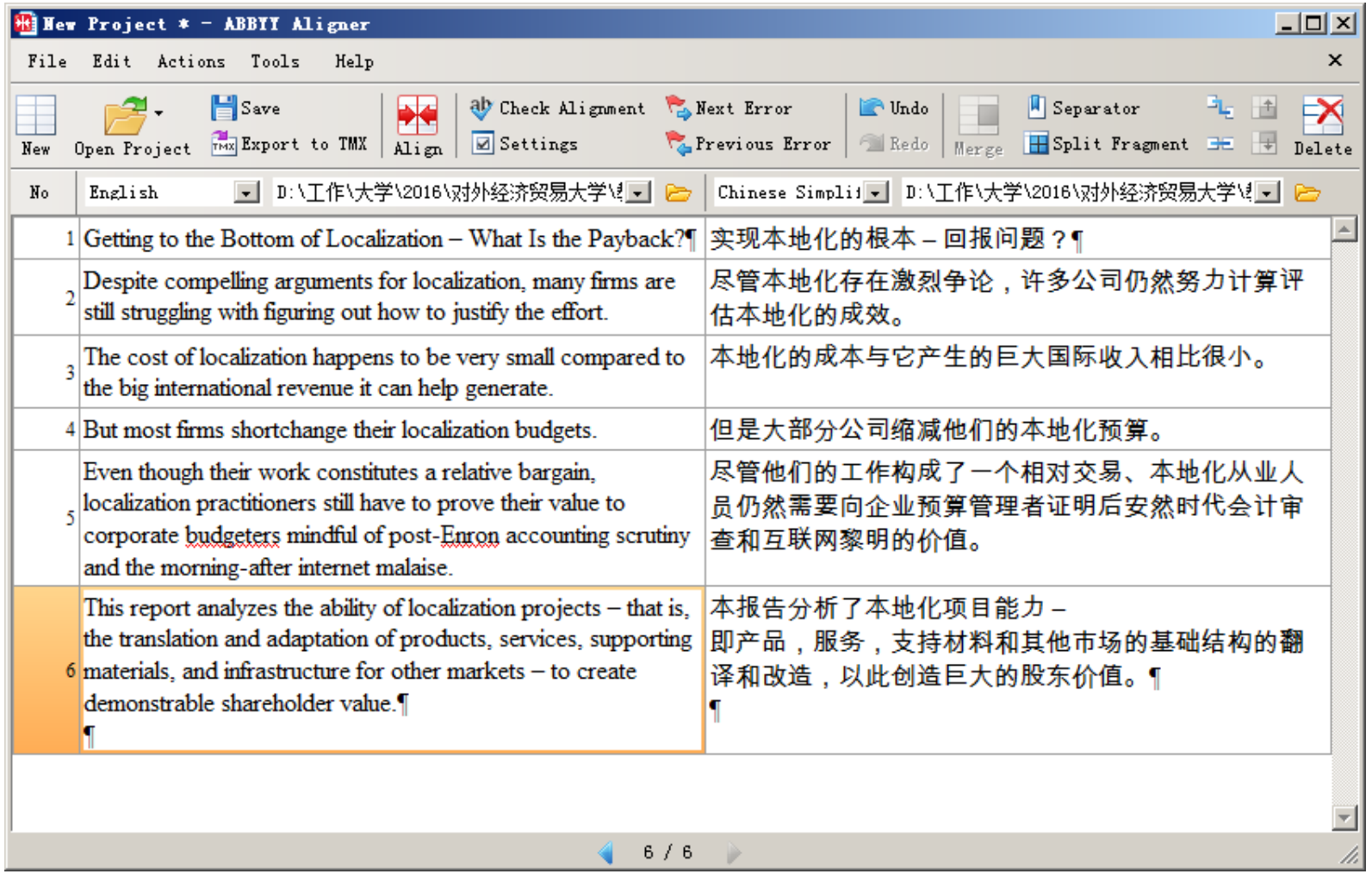
1. Download the tagger package for your system (PC-Linux, Mac OS-X (Intel), Mac OS-X (M1), ARM64, ARMHF, ARM-Android, PPC64le-Linux).
   If you have problems with your Linux kernel version, download this older Linux version and rename it to tree-tagger-linux-3.2.5.tar.gz.
2. Download the tagging scripts into the same directory.
3. Download the installation script install-tagger.sh.
4. Download the parameter files for the languages you want to process.
5. Open a terminal window and run the installation script in the directory where you have downloaded the files:
   sh install-tagger.sh
6. Make a test, e.g.
   echo 'Hello world!' | cmd/tree-tagger-english
   or
   echo 'Das ist ein Test.' | cmd/tagger-chunker-german
7. You also might want to have a look at my new part-of-speech tagger RNNTagger.

6. 语料库对齐工具

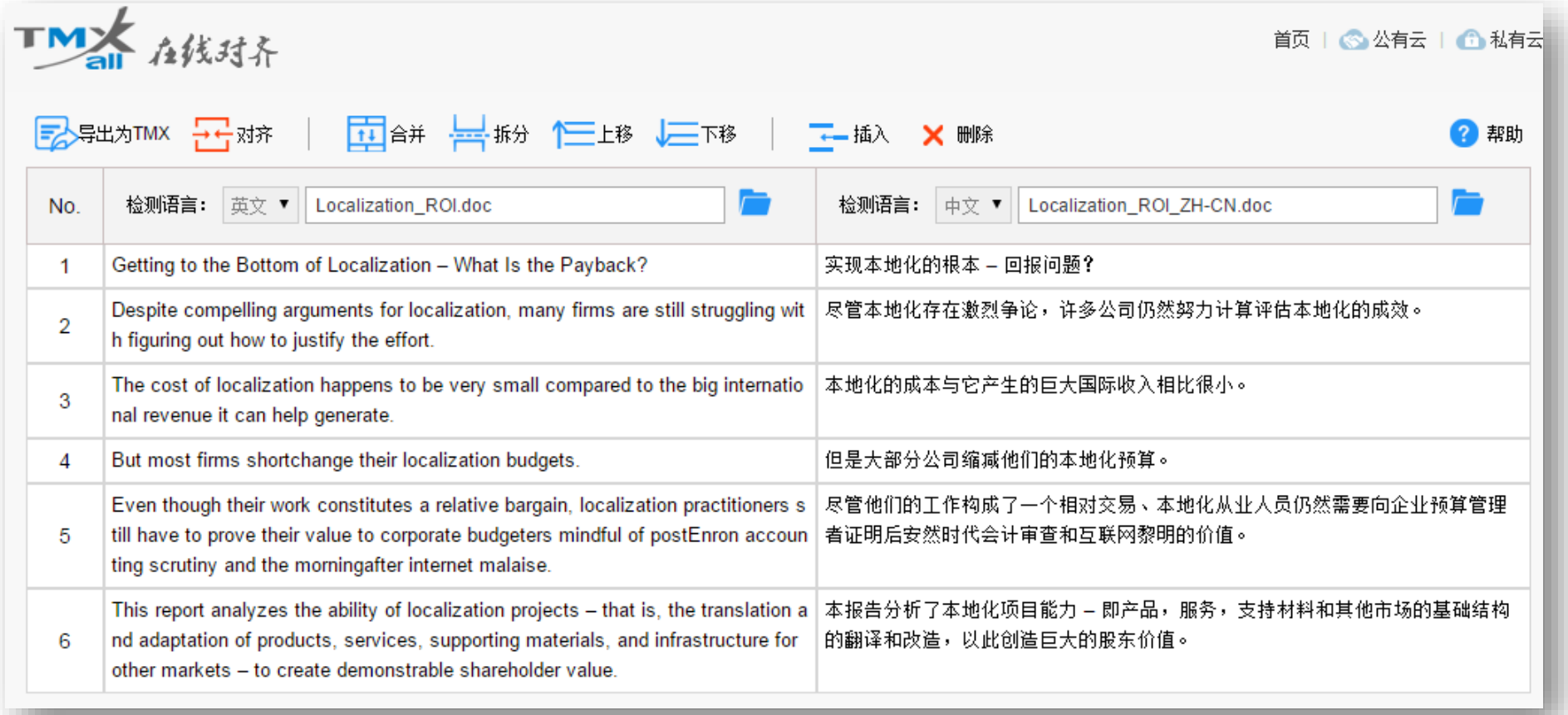
语料对齐

- **什么是对齐（Alignment）？**
  - 通过比较和关联源语言文档和目标语言文档创建双语平行语料库的过程。
- **对齐方法**
  - 使用CAT工具的对齐功能
  - 使用特定工具（Abbyy Aligner，TMXmall的在线对齐）
- **对齐注意事项**
  - 原文和译文段落数相同
  - 手工处理断句问题

使用Abbyy Aligner对齐



使用TMXMall在线对齐

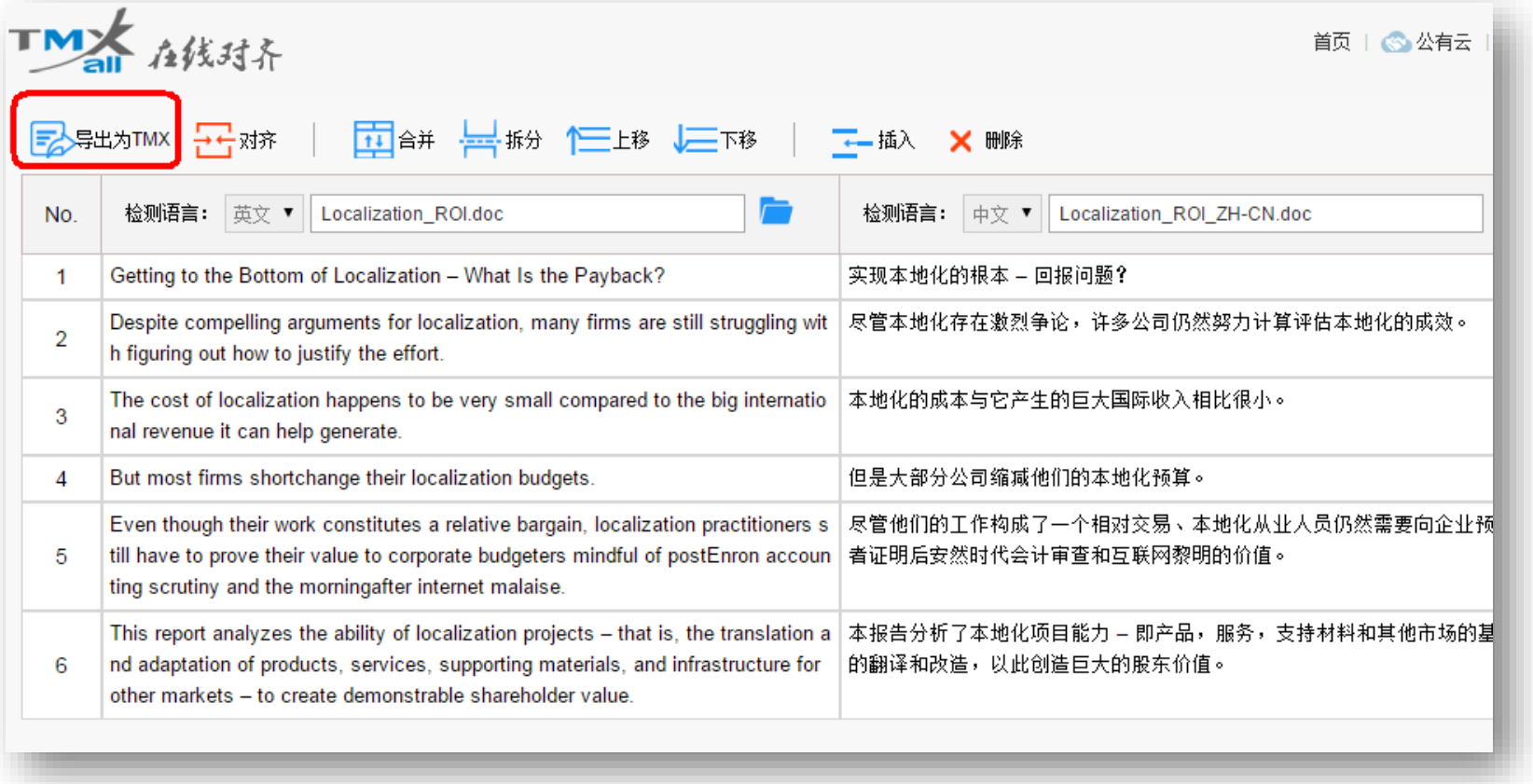


http://www.tmxmall.com

7. 翻译记忆库的创建和导出

翻译记忆库的导出

- **从TMXMall在线对齐工具导出**



ChatGPT创建翻译记忆库

将英文和中文以句子为单位对齐成标准TMX格式的翻译记忆库文件，下面第一段是英文原文，第二段是中文译文：

Getting to the Bottom of Localization – What Is the Payback?

Despite compelling arguments for localization, many firms are still struggling with figuring out how to justify the effort. The cost of localization happens to be very small compared to the big international revenue it can help generate. But most firms shortchange their localization budgets. Even though their work constitutes a relative bargain, localization practitioners still have to prove their value to corporate budgeters mindful of post-Enron accounting scrutiny and the morning-after internet malaise. This report analyzes the ability of localization projects – that is, the translation and adaptation of products, services, supporting materials, and infrastructure for other markets – to create demonstrable shareholder value.

实现本地化的根本－回报问题?

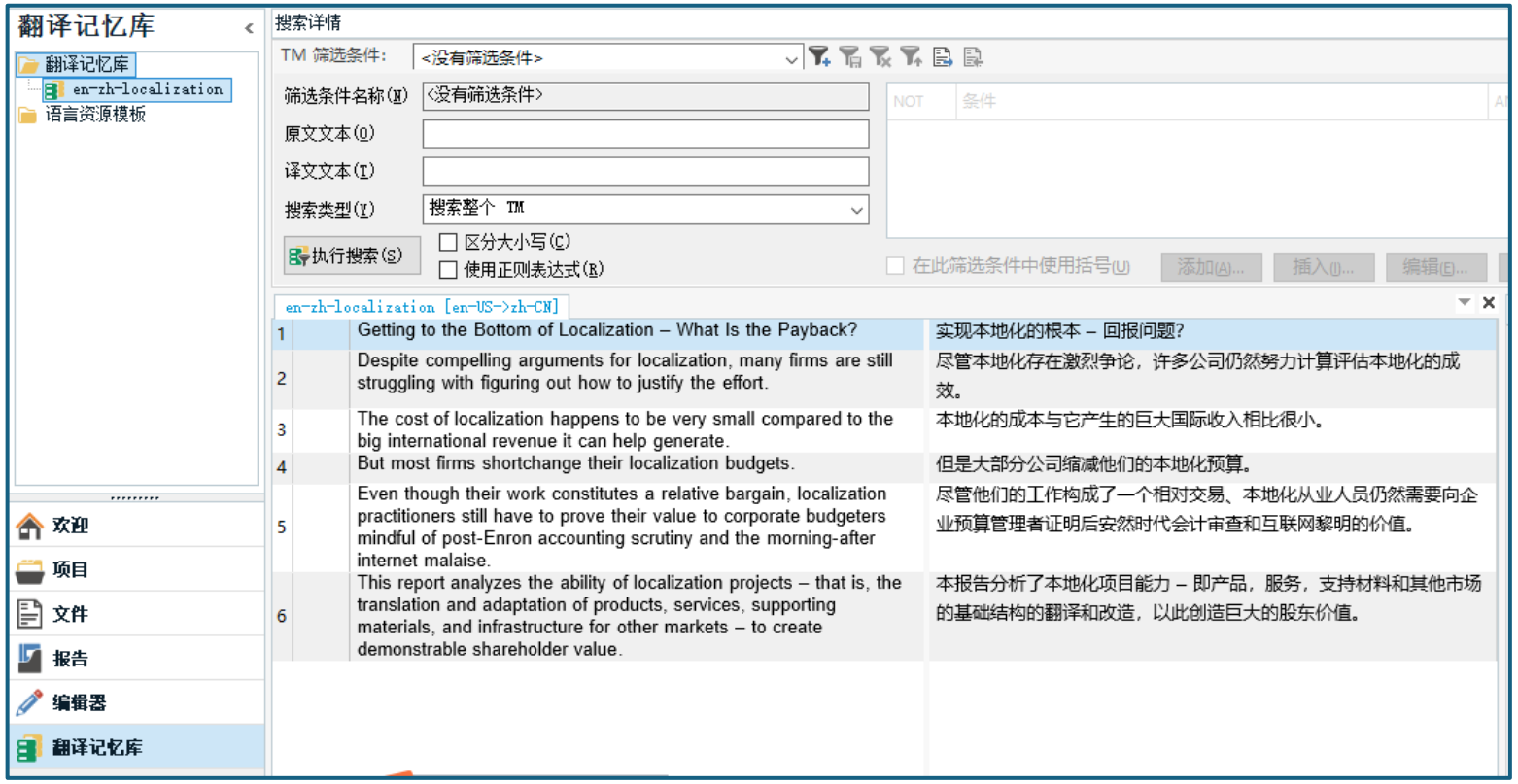
尽管本地化存在激烈争论，许多公司仍然努力计算评估本地化的成效。本地化的成本与它产生的巨大国际收入相比很小。但是大部分公司缩减他们的本地化预算。尽管他们的工作构成了一个相对交易、本地化从业人员仍然需要向企业预算管理者证明后安然时代会计审查和互联网黎明的价值。本报告分析了本地化项目能力 – 即产品，服务，支持材料和其他市场的基础结构的翻译和改造，以此创造巨大的股东价值。

ChatGPT创建翻译记忆库

<tmx version="1.4"> <header creationtool="Translation Memory eXchange" creationtoolversion="1.0" segtype="sentence" o-tmf="ABCTransMem" adminlang="en-US" srclang="en-US" datatype="PlainText" /> <body> <tu> <tuv xml:lang="en-US"><seg>Getting to the Bottom of Localization – What Is the Payback?</seg></tuv> <tuv xml:lang="zh-CN"><seg>实现本地化的根本 – 回报问题？</seg></tuv> </tu> <tu> <tuv xml:lang="en-US"><seg>Despite compelling arguments for localization, many firms are still struggling with figuring out how to justify the effort.</seg></tuv> <tuv xml:lang="zh-CN"><seg>尽管本地化存在激烈争论，许多公司仍然努力计算评估本地化的成效。</seg></tuv> </tu> <tu> <tuv xml:lang="en-US"><seg>The cost of localization happens to be very small compared to the big international revenue it can help generate.</seg></tuv> <tuv xml:lang="zh-CN"><seg>本地化的成本与它产生的巨大国际收入相比很小。</seg></tuv> </tu> <tu> <tuv xml:lang="en-US"><seg>But most firms shortchange their localization budgets.</seg></tuv> <tuv xml:lang="zh-CN"><seg>但是大部分公司缩减他们的本地化预算。</seg></tuv> </tu> <tu> <tuv xml:lang="en-US"><seg>Even though their work constitutes a relative bargain, localization practitioners still have to prove their value to corporate budgeters mindful of post-Enron accounting scrutiny and the morning-after internet malaise.</seg></tuv> <tuv xml:lang="zh-CN"><seg>尽管他们的工作构成了一个相对交易、本地化从业人员仍然需要向企业预算管理者证明后安然时代会计审查和互联网黎明的价值。</seg></tuv> </tu> <tu> <tuv xml:lang="en-US"><seg>This report analyzes the ability of localization projects – that is, the translation and adaptation of products, services, supporting materials, and infrastructure for other markets – to create demonstrable shareholder value.</seg></tuv> <tuv xml:lang="zh-CN"><seg>本报告分析了本地化项目能力 – 即产品，服务，支持材料和其他市场的基础结构的翻译和改造，以此创造巨大的股东价值。</seg></tuv> </tu> </body> </tmx>

ChatGPT创建翻译记忆库

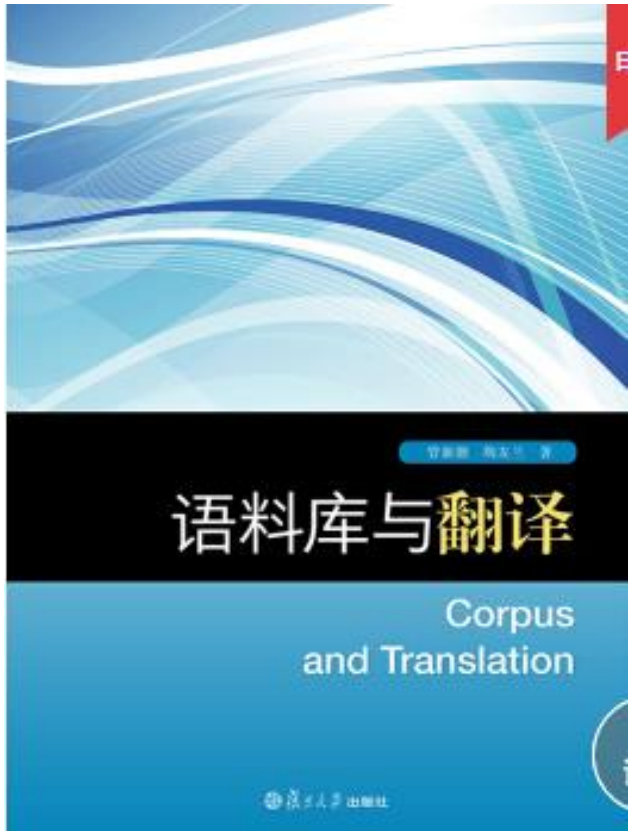
将文本内容保存到TXT文件，修改文件扩展名为TMX，在Trados Studio新建翻译记忆库文件，导入TMX文件。

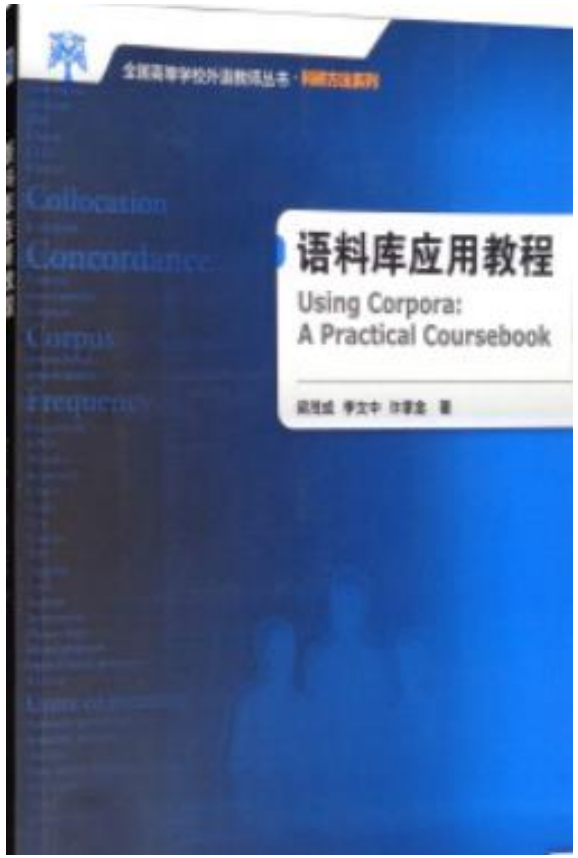


课外阅读材料

Excel文件与TMX文件的相互转换方法

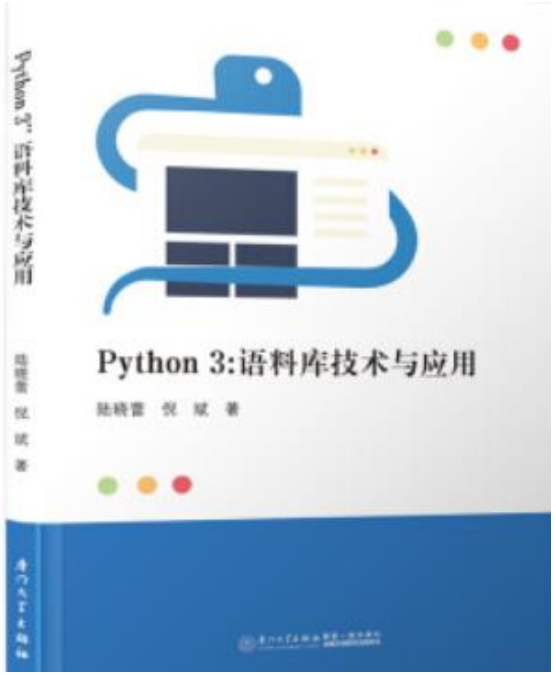
来源：“本地化世界”微信公众号

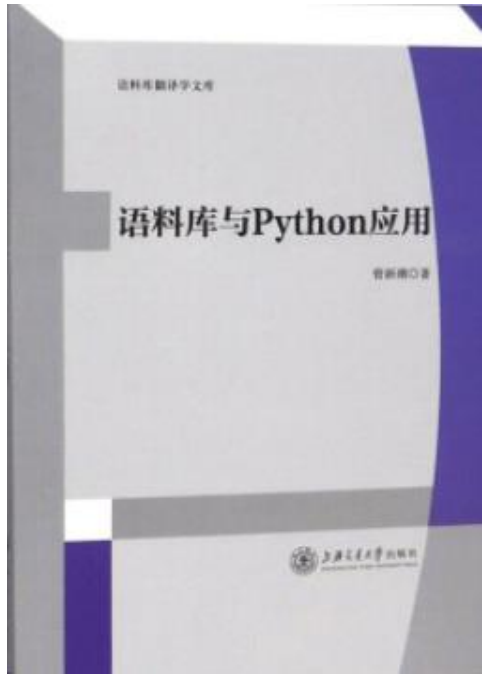




参考材料

Python语料库编程

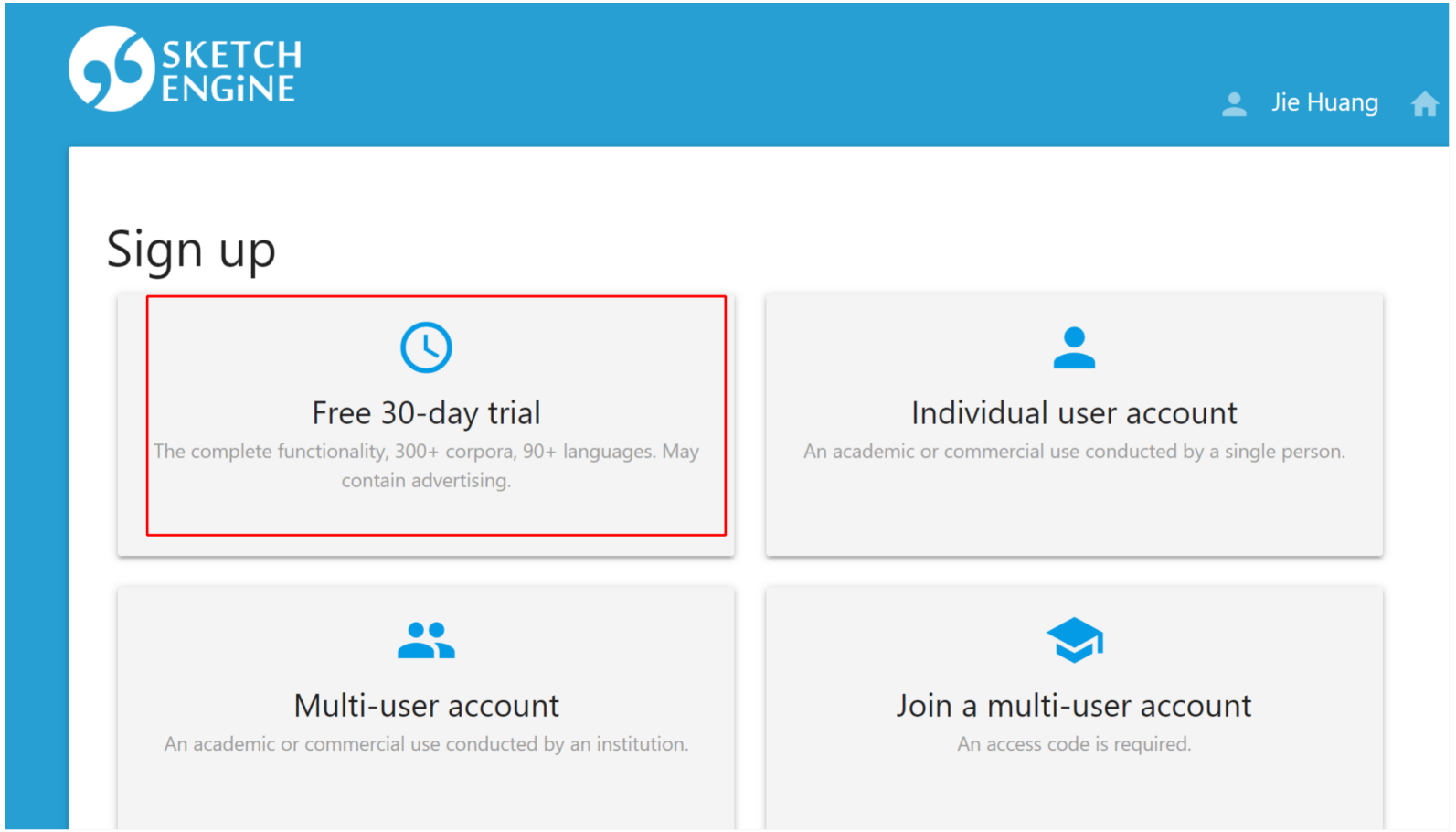




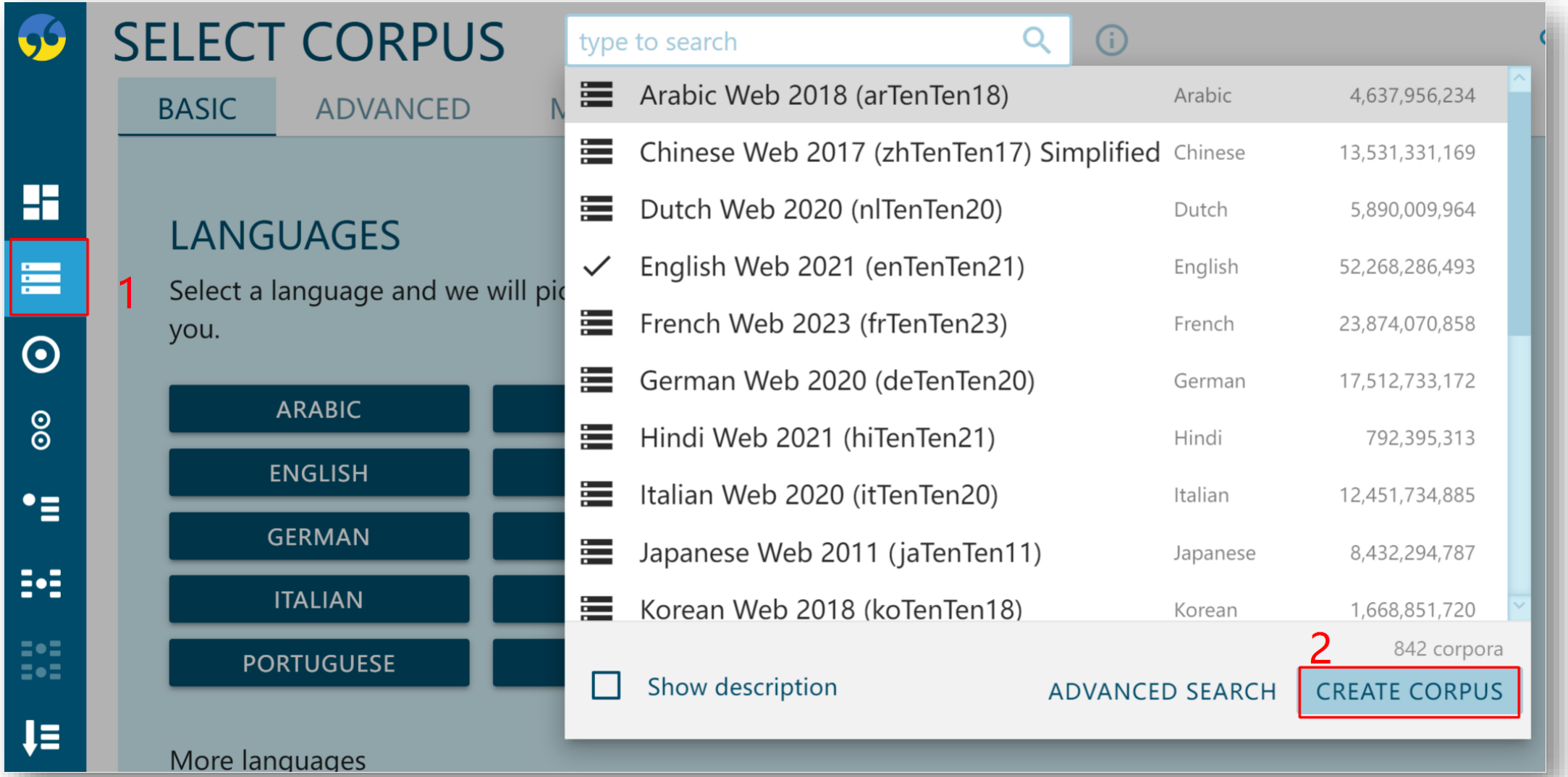
番外：SketchEngine 注册和使用指南

注册 30-day free trial

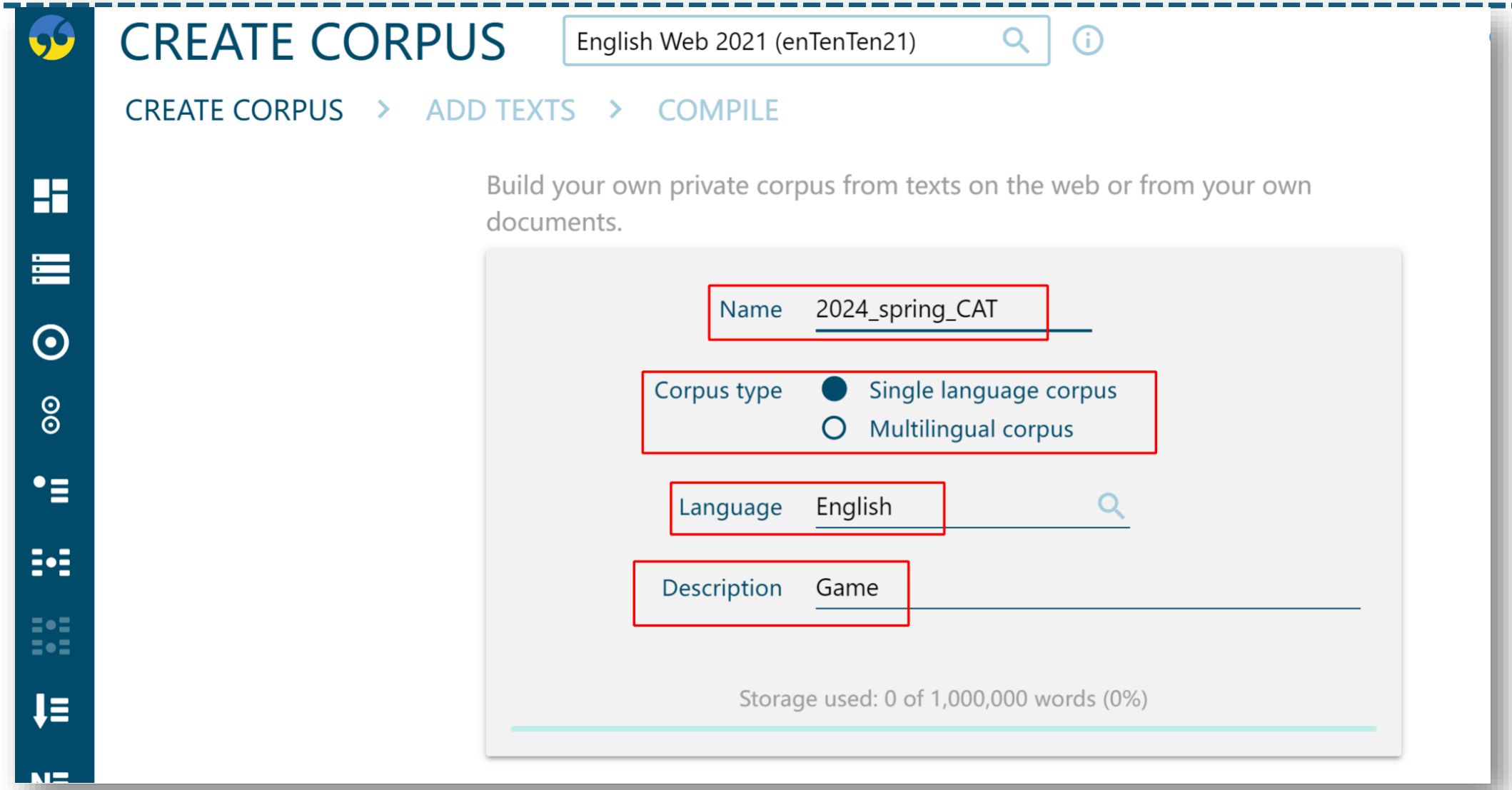
https://auth.sketchengine.eu/#register



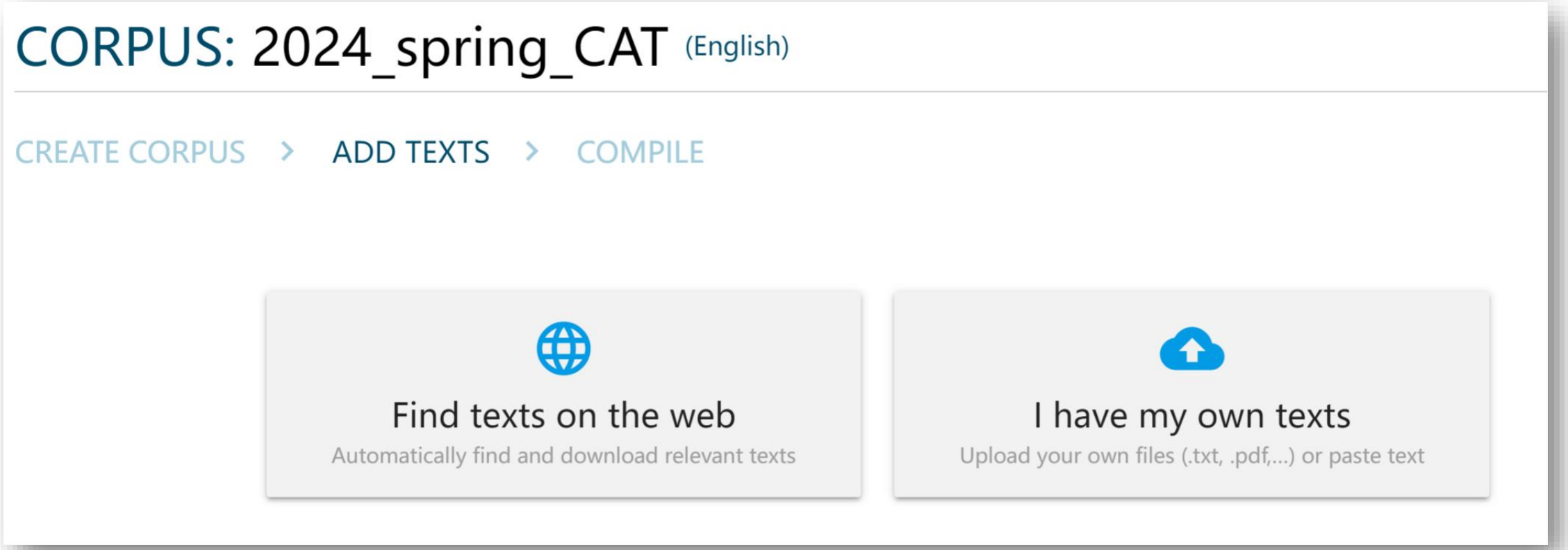
创建语料库—create corpus



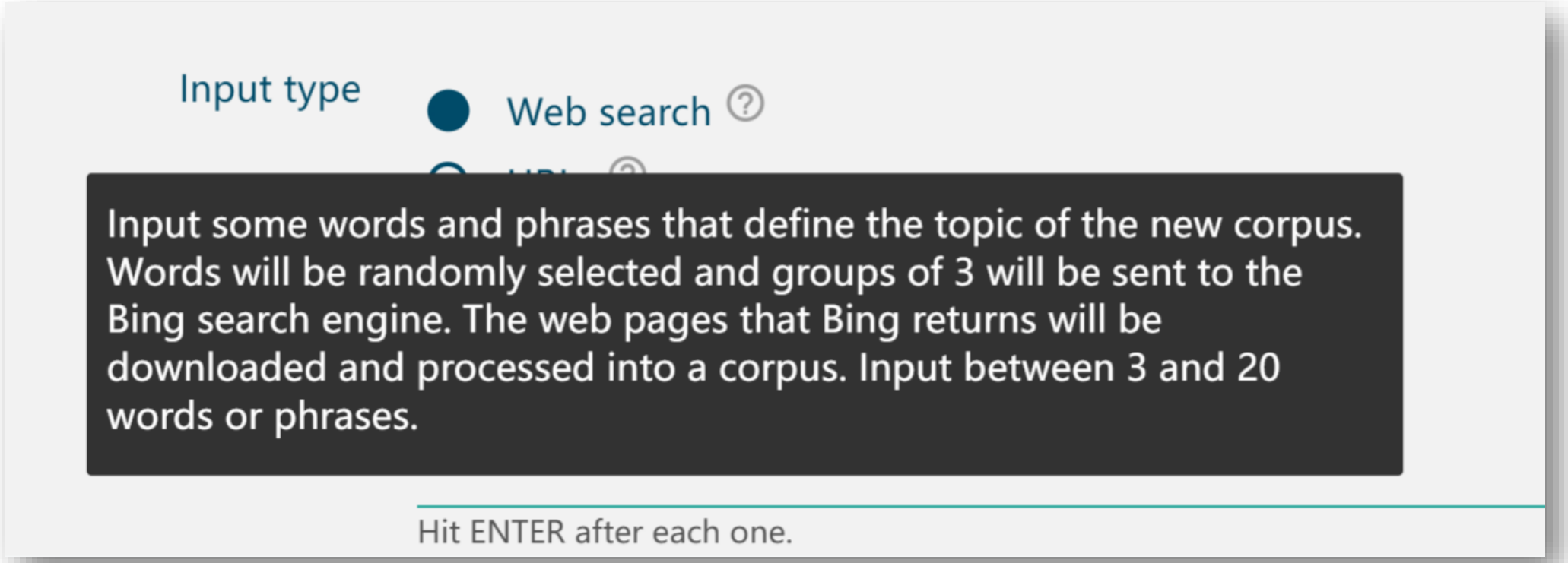
设置语料库属性


CREATE CORPUS
English Web 2021 (enTenTen21)
CREATE CORPUS
ADD TEXTS
COMPILE
Build your own private corpus from texts on the web or from your own documents.
Name
2024_spring_CAT
Corpus type
Single language corpus
Multilingual corpus
Language
English
Description
Game
Storage used: 0 of 1,000,000 words (0%)

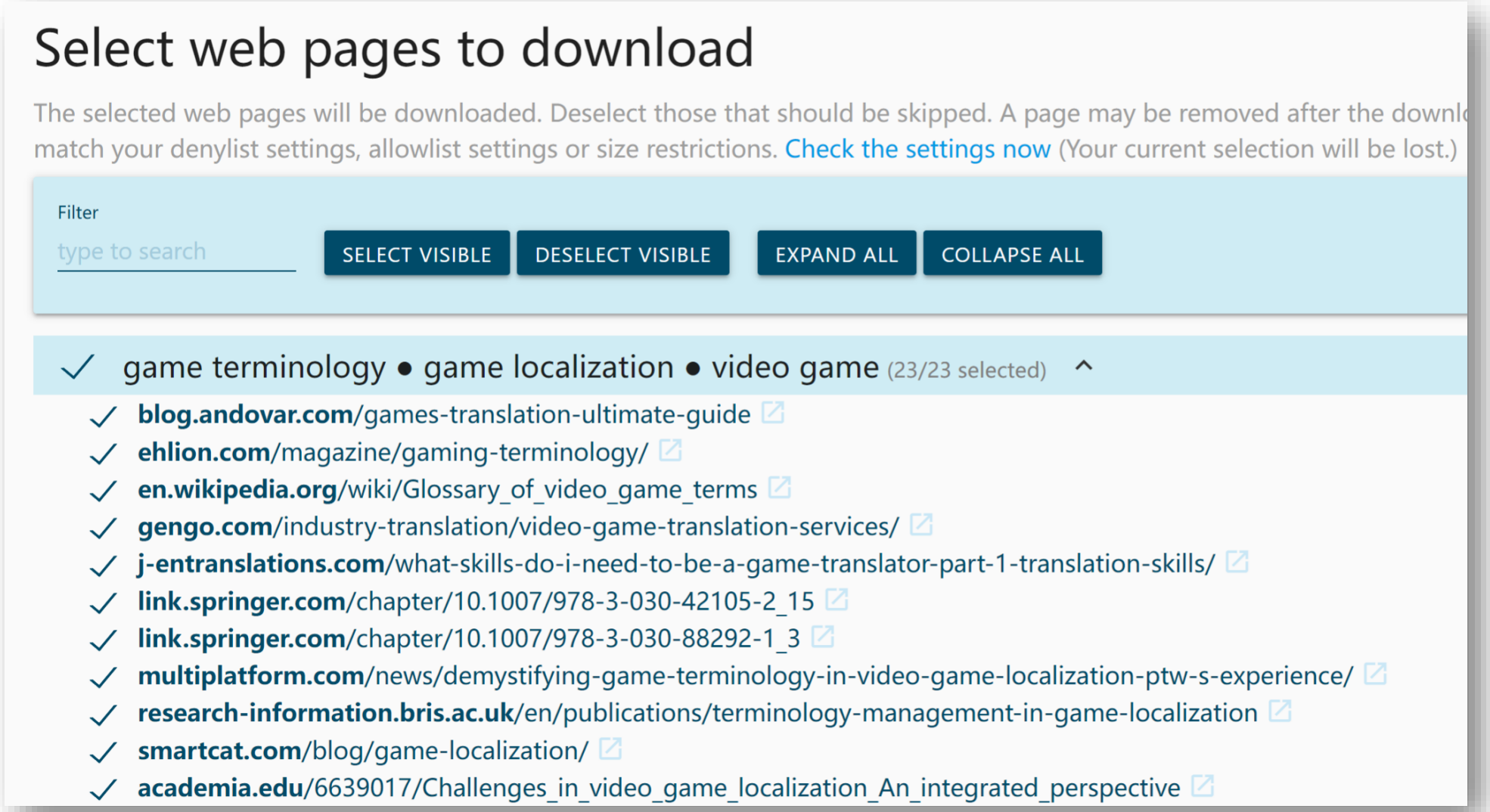
选择语料来源——网页/本地文件



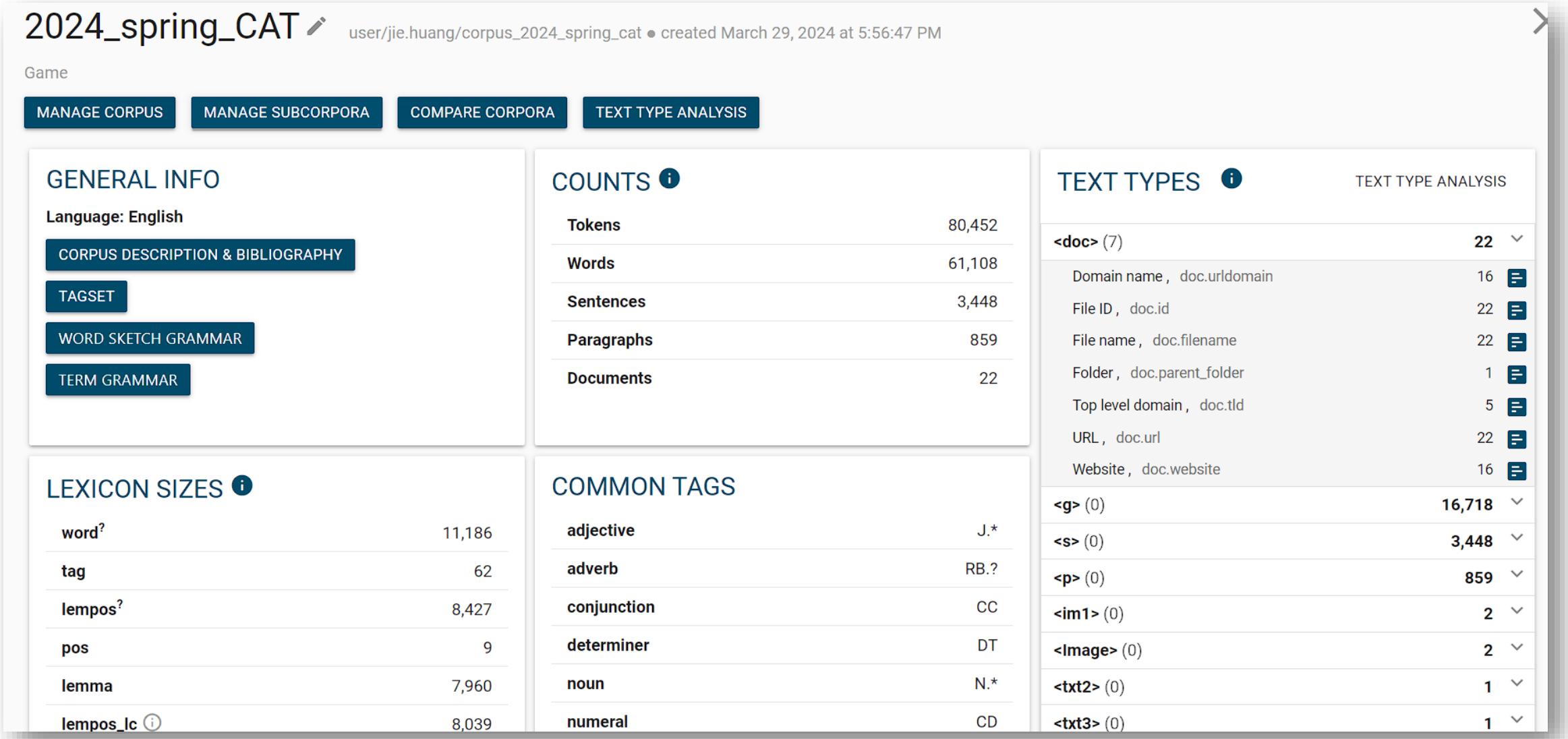
网页搜索关键词：至少3个



搜索——GO


Select web pages to download
The selected web pages will be downloaded. Deselect those that should be skipped. A page may be removed after the downl
match your denylist settings, allowlist settings or size restrictions. Check the settings now (Your current selection will be lost.)
Filter
type to search
SELECT VISIBLE
DESELECT VISIBLE
EXPAND ALL
COLLAPSE ALL
game terminology • game localization • video game (23/23 selected)
blog.andovar.com/games-translation-ultimate-guide
ehlion.com/magazine/gaming-terminology/
en.wikipedia.org/wiki/Glossary_of_video_game_terms
gengo.com/industry-translation/video-game-translation-services/
j-entranslations.com/what-skills-do-i-need-to-be-a-game-translator-part-1-translation-skills/
link.springer.com/chapter/10.1007/978-3-030-42105-2_15
link.springer.com/chapter/10.1007/978-3-030-88292-1_3
multiplatform.com/news/demystifying-game-terminology-in-video-game-localization-ptw-s-experience/
research-information.bris.ac.uk/en/publications/terminology-management-in-game-localization
smartcat.com/blog/game-localization/
academia.edu/6639017/Challenges_in_video_game_localization_An_integrated_perspective

语料库创建成功


2024_spring_CAT
user/jie.huang/corpus_2024_spring_cat • created March 29, 2024 at 5:56:47 PM
Game
MANAGE CORPUS
MANAGE SUBCORPORA
COMPARE CORPORA
TEXT TYPE ANALYSIS
GENERAL INFO
Language: English
CORPUS DESCRIPTION & BIBLIOGRAPHY
TAGSET
WORD SKETCH GRAMMAR
TERM GRAMMAR
COUNTS
Tokens 80,452
Words 61,108
Sentences 3,448
Paragraphs 859
Documents 22
TEXT TYPES
TEXT TYPE ANALYSIS
<doc> (7) 22
Domain name , doc.urldomain 16
File ID , doc.id 22
File name , doc.filename 22
Folder , doc.parent_folder 1
Top level domain , doc.tld 5
URL , doc.url 22
Website , doc.website 16
<g> (0) 16,718
<s> (0) 3,448
<p> (0) 859
<im1> (0) 2
<Image> (0) 2
<txt2> (0) 1
<txt3> (0) 1
LEXICON SIZES
word? 11,186
tag 62
lempos? 8,427
pos 9
lemma 7,960
lempos_lc 8,039
COMMON TAGS
adjective J.*
adverb RB.?
conjunction CC
determiner DT
noun N.*
numeral CD

术语列表

KEYWORDS 2024_spring_CAT

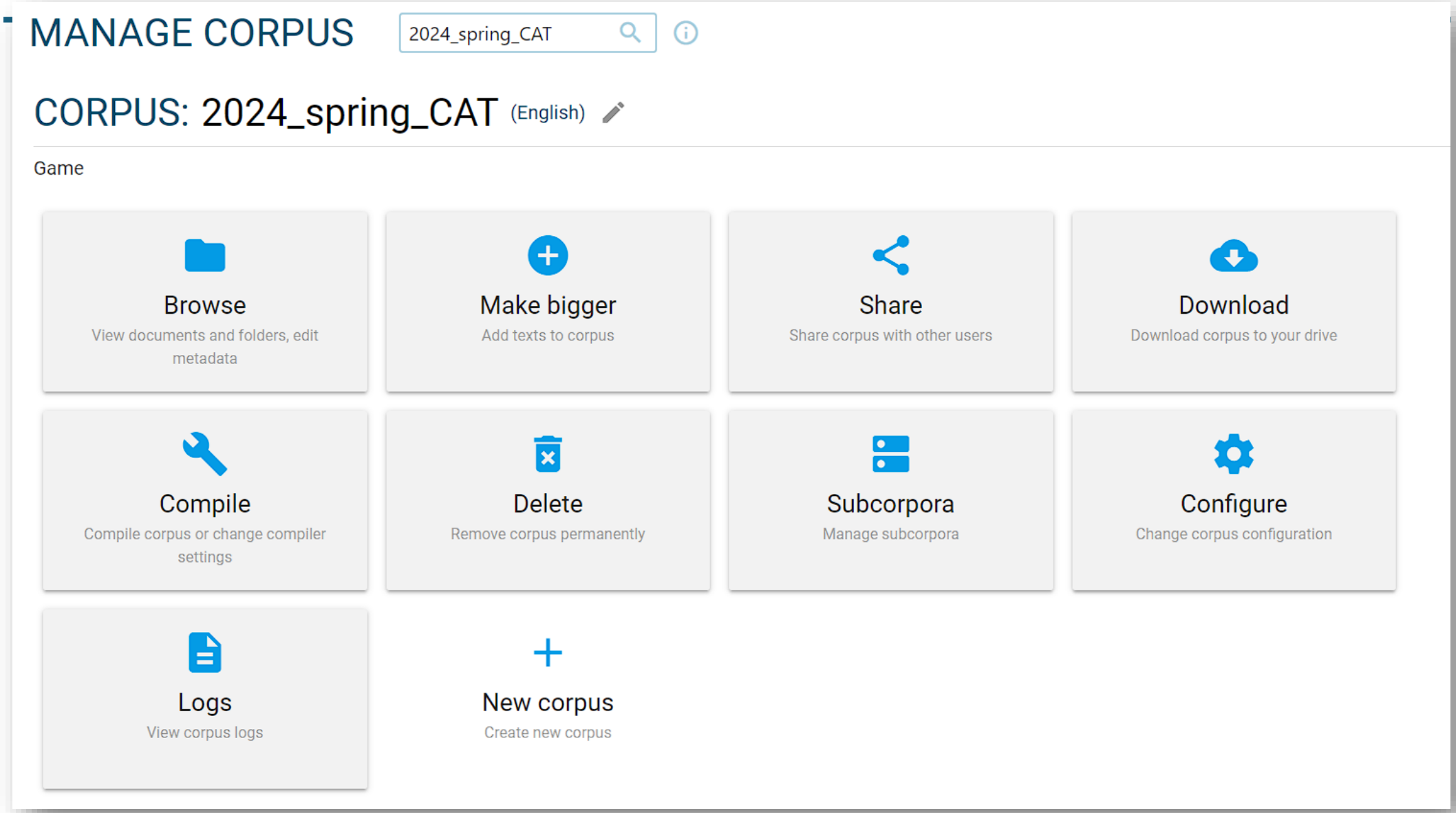
SINGLE-WORDS ✓ MULTI-WORD TERMS ✓

reference corpus: English Web 2021 (enTenTen21) (items: 7,271)

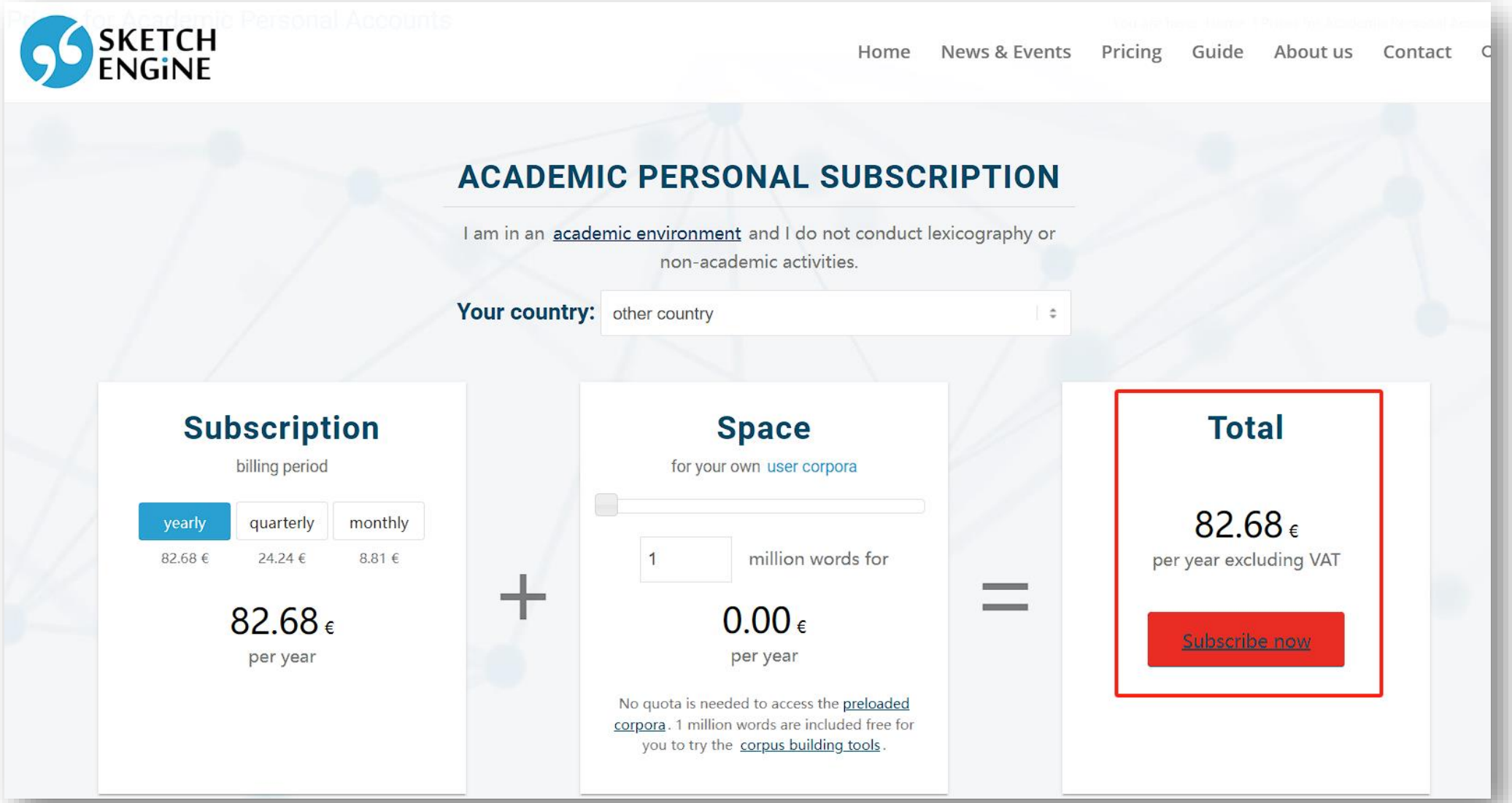
	Lemma		Lemma		Lemma		Lemma		Lemma
1	llm	11	openai	21	generative	31	instructibility	41	eva-clip
2	llms	12	multi-modal	22	ai-native	32	vision-language	42	xu
3	lmms	13	llava	23	pre-training	33	cvf	43	chatbot
4	arxiv	14	pre-trained	24	llm-based	34	cogvlm	44	human-like
5	lmm	15	modality	25	gpt-4v	35	rlhf	45	llama
6	chatgpt	16	peft	26	imagebind	36	fine-tuning	46	zhao
7	gpt-4	17	sft	27	vit	37	blip-2	47	neuro-symbolic
8	multimodal	18	mm-llm	28	next-gpt	38	image-text	48	encoder
9	preprint	19	gpt-3	29	q-former	39	vicuna	49	zhang
10	mm-llms	20	models	30	webvoyager	40	instructblip	50	tangibility

Rows per page: 50 1–50 of 100 1 / 2

语料库管理


MANAGE CORPUS
2024_spring_CAT
CORPUS: 2024_spring_CAT (English)
Game
Browse
View documents and folders, edit metadata
Make bigger
Add texts to corpus
Share
Share corpus with other users
Download
Download corpus to your drive
Compile
Compile corpus or change compiler settings
Delete
Remove corpus permanently
Subcorpora
Manage subcorpora
Configure
Change corpus configuration
Logs
View corpus logs
New corpus
Create new corpus

Pricing


SKETCH ENGiNE
Home
News & Events
Pricing
Guide
About us
Contact
ACADEMIC PERSONAL SUBSCRIPTION
I am in an academic environment and I do not conduct lexicography or non-academic activities.
Your country:
other country
Subscription
billing period
yearly
quarterly
monthly
82.68 €
24.24 €
8.81 €
82.68 €
per year
+
Space
for your own user corpora
1
million words for
0.00 €
per year
No quota is needed to access the preloaded corpora. 1 million words are included free for you to try the corpus building tools.
=
Total
82.68 €
per year excluding VAT
Subscribe now

Sketch engine vs. English-corpora

english-corpora.org/compare-sketchEngine.asp

English-Corpora.org

corpora guides related resources users my account upgrade help

English-Corpora.org and SketchEngine are probably the two largest sites for online corpora. We believe that both sites provide valuable resources for linguists, lexicographers, and language learners and teachers.

The following is a comparison of the two sites, for those who are already family with Sketch Engine, but are new to English-Corpora.org. Admittedly (because this list is at English-Corpora.org), it is probably biased towards English-Corpora.org, and we invite you to look more in depth at what Sketch Engine has to offer as well. Finally, if there is incomplete / incorrect information below, please let us know.

Feature	Sketch Engine	English-Corpora.org
Corpora	- Extremely wide (90+) range of languages, and hundreds of corpora - For English, very large web-based corpora, as well as many other specialized corpora	• Mostly English, as well as some for Spanish and Portuguese • For English, perhaps the best suite of corpora for looking at variation: genre-based, historical, and dialectal • Largest corpora are iWeb (14 billion words) and NOW (14.6 billion words and growing by ~250 million words each month)
Users / research	- Linguistics and lexicographers, teachers and learners, etc (For those with information on Sketch Engine, please send us more detailed / verifiable information on number of users, researchers, universities with licenses, number of publications, etc)	- ~130,000 distinct users each month, including about 80,000 registered users - ~300 universities have academic (group) licenses, as well as large government-funded licenses - More than 16,500 registered "researchers" (professors or graduate students) in linguistics or language studies - Cited in more than 10,000 academic publications, including more than 5,000 in the past five years - The data (e.g. full-text, word frequency) is used by hundreds of companies, including Google, Amazon, Microsoft, IBM, Samsung; Merriam-

End