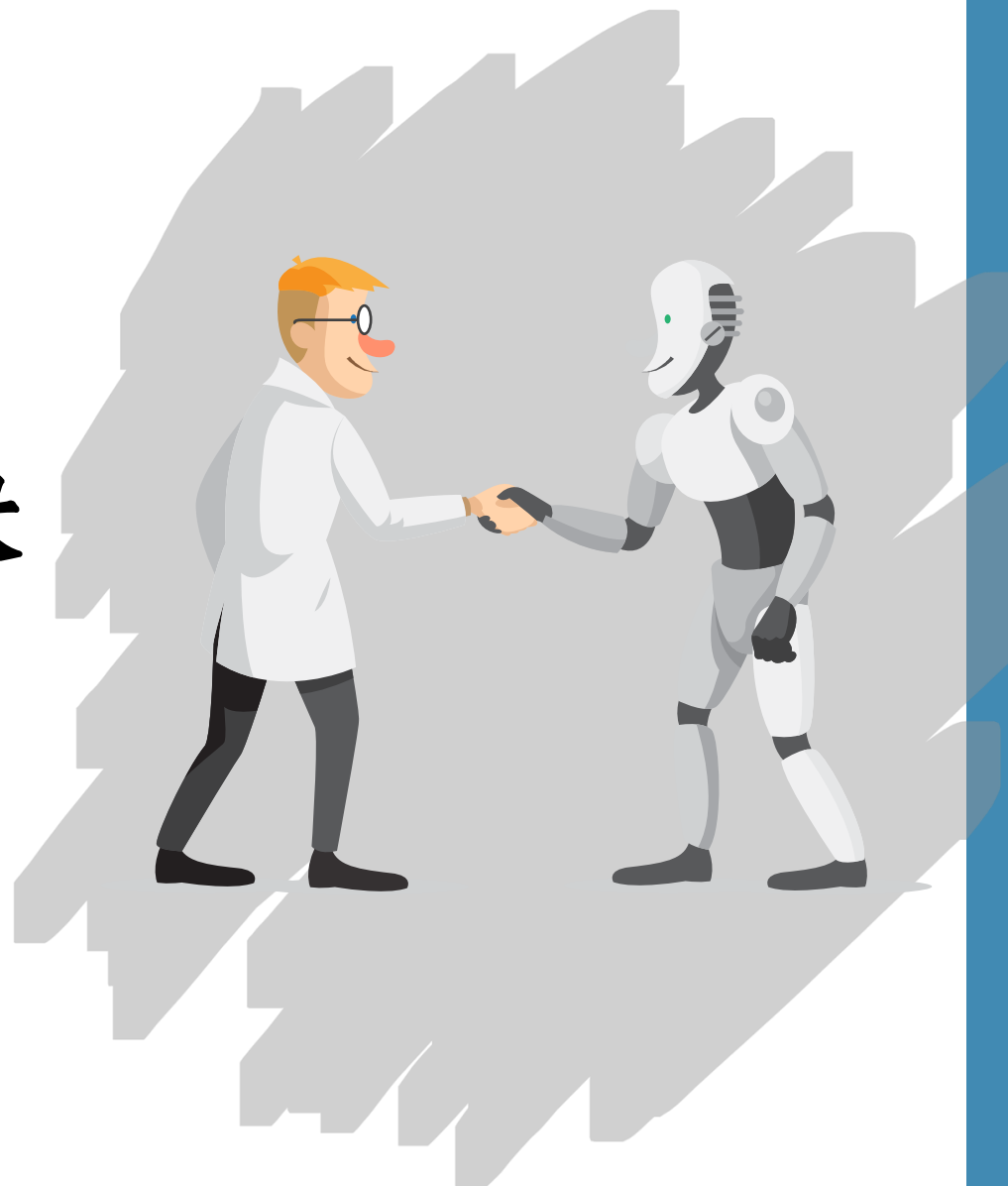


第3讲：文件格式转换

黄婕

2025年3月20日



本节内容



1. 文件格式转换：
需求和目标

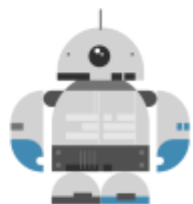
2. OCR 使用场景和
工具



3. PDF文件的创建、
转换及翻译

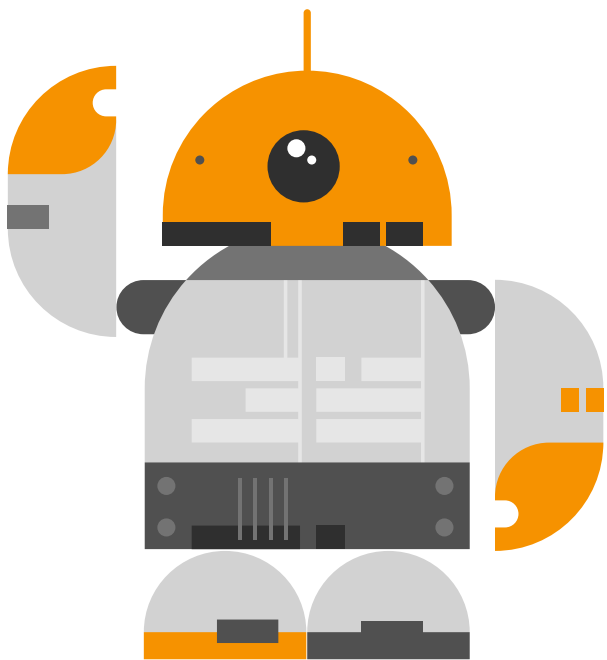


4. 案例练习



5. 小组作业





1.文件格式转换: 需求和目标

假如你是翻译项目经理，遇到以下任务.....

你需要至少完成几步操作？

营销宣传册翻译

- 一家国际企业需要将其产品宣传册翻译成多种语言。
- 原始文件是 InDesign 格式（.indd）

电子书出版多语言版本

- 一家出版商希望翻译一本电子书，以进入新的语言市场。
- 原始文件是电子书格式（.epub）

法律合同翻译

- 翻译一个法律合同，其中包含复杂的文本和格式要求
- 原始文档以PDF格式提供

软件用户界面翻译

- 一个软件开发公司需要本地化其软件的用户界面
- 包含大量XML格式的字符串和标签

多个项目的翻译记忆更新

- 一名翻译项目经理需要整合来自多语言项目的翻译记忆，以便日后使用。
- 文件格式TMX

为什么要格式转换？

处理多样性文件格式

- 转换为**可编辑和可处理**的格式，可直接在CAT工具中进行文本处理和翻译，提高效率和准确性。

保留文件格式和布局

- 在翻译前确保从原始文件中提取文本，同时**在翻译后恢复原有的格式和样式**。这在客户要求严格保留排版和设计的项目中尤为重要（法律、营销、软件等）。

术语和内容的批量处理（长久之计）

- 预处理文件以提取术语库和翻译记忆（TM）是一个关键步骤，可提高**术语一致性**，还能利用之前的**翻译记忆**，提高翻译效率和质量。

常见文件格式

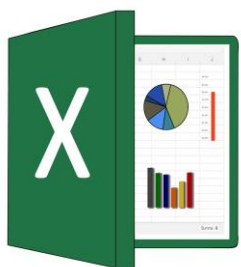


文本文件 (.txt):

- 纯文本文件是最简单和最常见的格式，可以包含简单的文本信息，例如程序代码、配置文件等。通常用于软件本地化中的资源文件。

标记语言文件 (.xml, .html):

- XML（可扩展标记语言）和HTML（超文本标记语言）文件常用于网页和软件界面的本地化，其中包含文本和标记，需要在翻译过程中保持标记结构的完整性。



电子表格文件 (.xlsx, .xls, .csv):

- 电子表格文件通常包含大量的数据和文本，用于本地化的文本可能包括产品名称、描述、价格等，尤其在电子商务领域的本地化中常见。

字幕文件 (.srt, .sub, .ass):

- 字幕文件用于电影、电视剧等视频内容的本地化，其中包含对话文本及相应的时间轴信息。



资源文件 (.resx, .po, .mo, .xliff):

- 资源文件是软件本地化中常见的格式，包含了软件界面、按钮标签、提示信息等文本信息，以便在不同语言环境下进行切换。

数据库文件 (.sql, .db):

- 数据库文件中包含了软件或网站的文本信息，用于动态生成页面内容或应用程序中的文本信息。



桌面出版文件 (.indd, .qxp):

- 桌面出版文件用于制作印刷品，包括书籍、杂志、宣传册等，其中包含排版、文本、图像等内容，需要进行本地化以适应不同的语言和文化环境。

多媒体文件 (.pdf, .docx, .pptx):

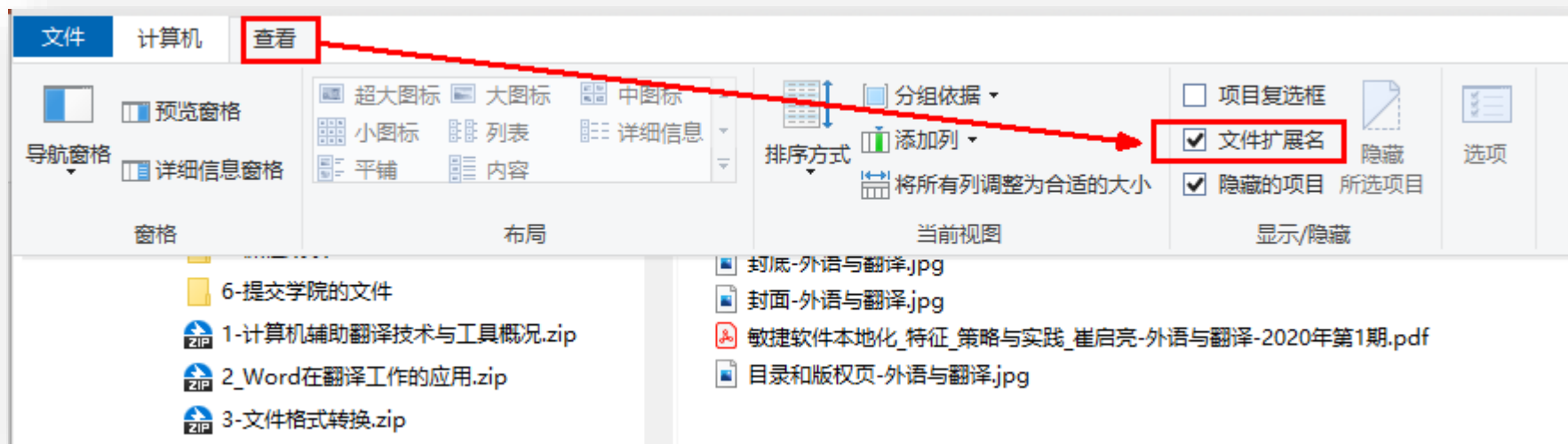
- 多媒体文件中可能包含了文本信息，例如PDF文件中的文本内容、Word文档中的说明、PowerPoint演示文稿中的文字等，需要进行本地化以确保内容的准确传达。



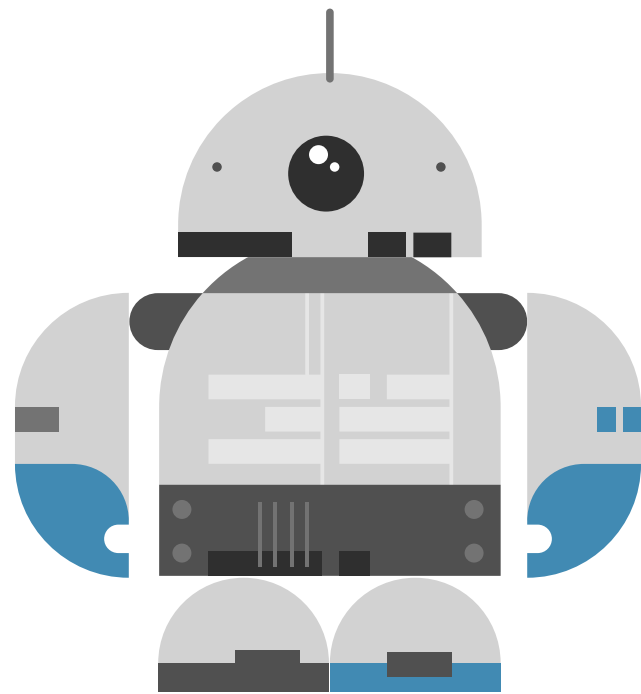
Acrobat Pro DC

初始：Windows显示文件扩展名的设置

- 文件类型通过文件扩展名显示和查看。
- Windows默认不显示文件的扩展名。
- 设置Windows 10显示文件扩展名
 1. 打开Windows资源管理器（快捷键：Win窗口键+E）
 2. “查看” -> 选中“文件扩展名”



2. OCR 使用场景和工具



OCR 光学字符识别

OCR 含义：Optical Character Recognition 光学字符识别

OCR 使用场景：

- **文档数字化**：将纸质文件转换为数字格式，以便于存储、检索和编辑。
- **翻译工作**：在翻译过程中，OCR可以帮助将书籍、报纸等纸质材料的文本提取出来，方便后续的翻译和编辑。
- **助力视障人士**：OCR技术可以将印刷文本转换为语音或盲文，帮助视障人士获取信息。
- **历史文献保存**：将古籍、手稿等历史文献数字化，便于保存和研究。
- **车牌识别**：在交通管理中，OCR用于识别和记录车辆的车牌信息。



文本信息
数字化

常用场景和工具

日常手机使用

CamScanner

手机拍照截图-文本识别

微信图片识别

日常电脑使用

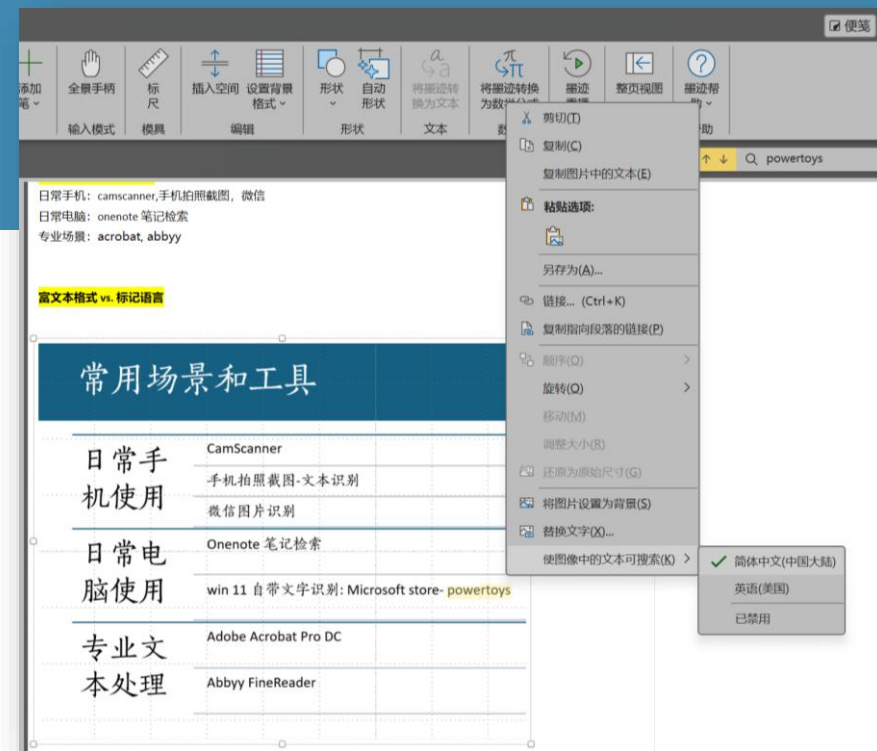
Onenote 笔记检索

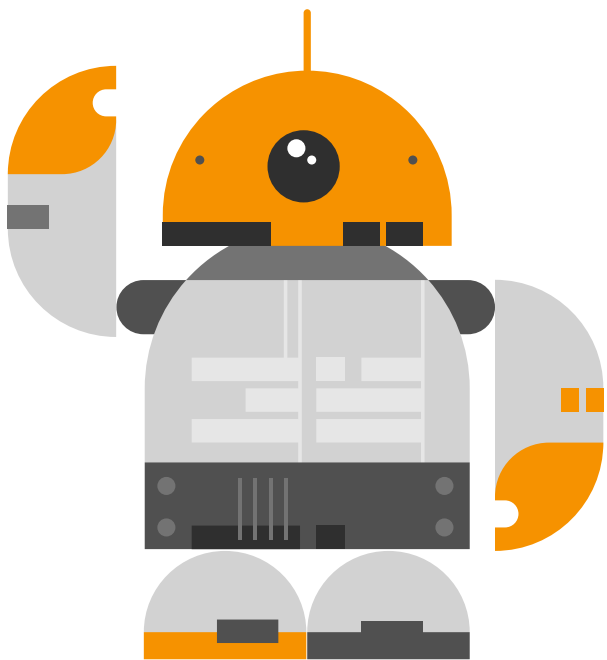
win 11 自带文字识别: Microsoft store-
powertoys – (Win + shift + T)

专业文本处理

Adobe Acrobat Pro DC

Abbyy FineReader





3. PDF文件的创建、 转换及翻译

翻译PDF文件

PDF全称**Portable Document Format**，便携式文档格式，是一种电子文件格式。

文件格式与操作系统平台无关，PDF文件不论在**Windows**，**Unix**还是在**Mac OS**操作系统中都是通用的。

成为在Internet上进行电子文档发行和数字化信息传播的理想文档格式。越来越多的电子图书、产品说明、公司文告、网络资料、电子邮件开始使用PDF格式文件。



创建PDF文件的工具

Microsoft Office 2010及以上版本

Adobe Acrobat Professional (DC)

Adobe FrameMaker

Adobe InDesign

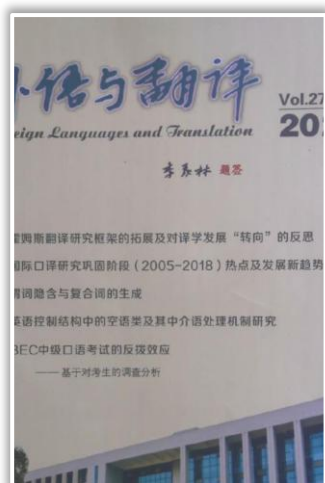
扫描仪 (Scanner)

PDF文件的类型及其转换方法



可编辑的PDF文件

- 特征：文本可选中，复制到Word后，可编辑
- 内容来源：由word, excel, ppt或InDesign, Framemaker 等软件生成的含有文本信息的PDF
- 常用转换工具：**Adobe Acrobat Pro DC**



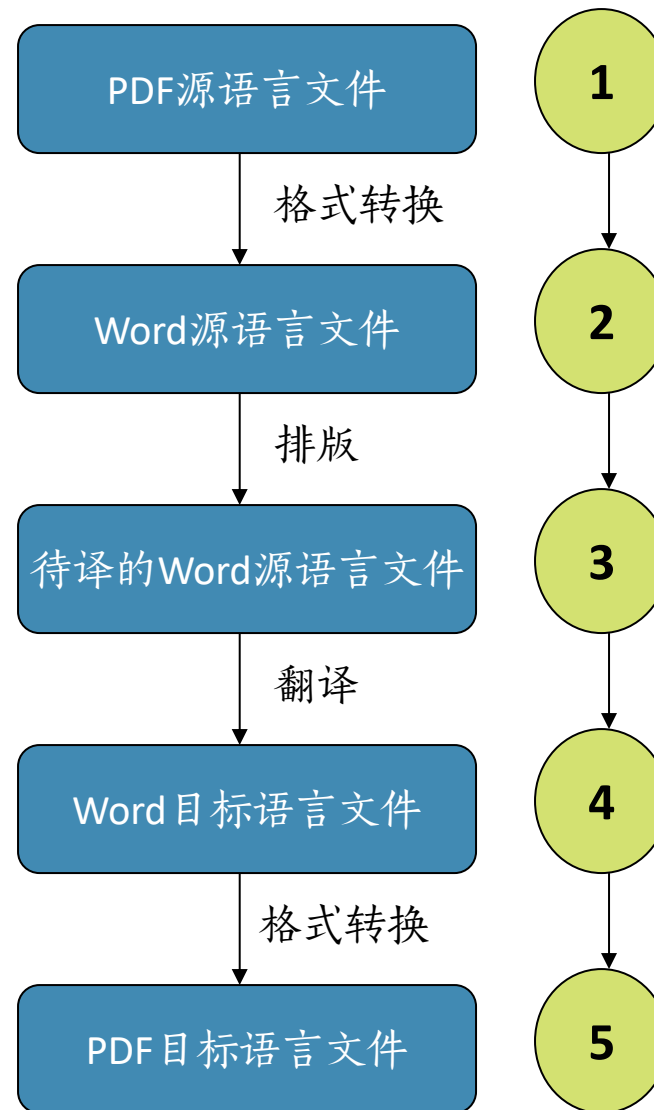
图片格式的PDF文件

- 内容来源：由图像或扫描文档转换而来
- 特征：选取内容复制到Word后，是图像格式，不可编辑
- 常用转换工具：**Abbyy FineReader**

翻译PDF文件的思路

最理想的翻译方法：
原始格式文件

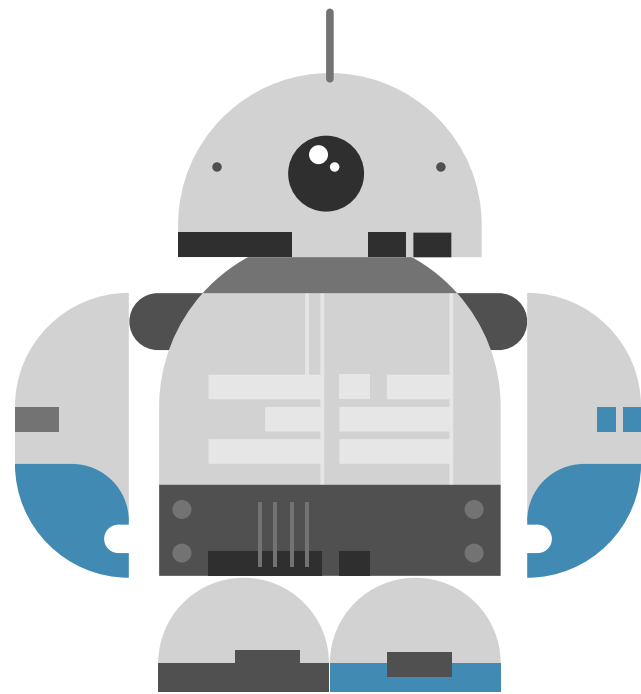
- Word格式
- InDesign的Indd格式
- Illustrator的AI格式



翻译PDF文件的具体步骤

1. 将PDF**转换**为Word文件(图片格式PDF使用FineReader; 非图片格式的PDF使用Acrobat Pro DC)
2. 对转换后的Word文件进行**修正并排版**
3. **翻译**Word文件 (包括图像处理)
4. 查看和调整翻译后的Word文件 (**排版**)
5. 将Word文件**转换**为PDF文件

4. 案例练习



文件管理思路

- 1_Source
- 2_Splitted
- 3_PreConvert
- 4_Pre-DTP
- 5_Translation
- 6_Post-Convert

1. Source 文件单独保存

- 原始文件备份

2. 各步骤文档留痕

- 以便纠正错误、减小影响

3. 文档命名

- 下划线、字母首字母大写区分

PDF编辑软件选择

Adobe Acrobat

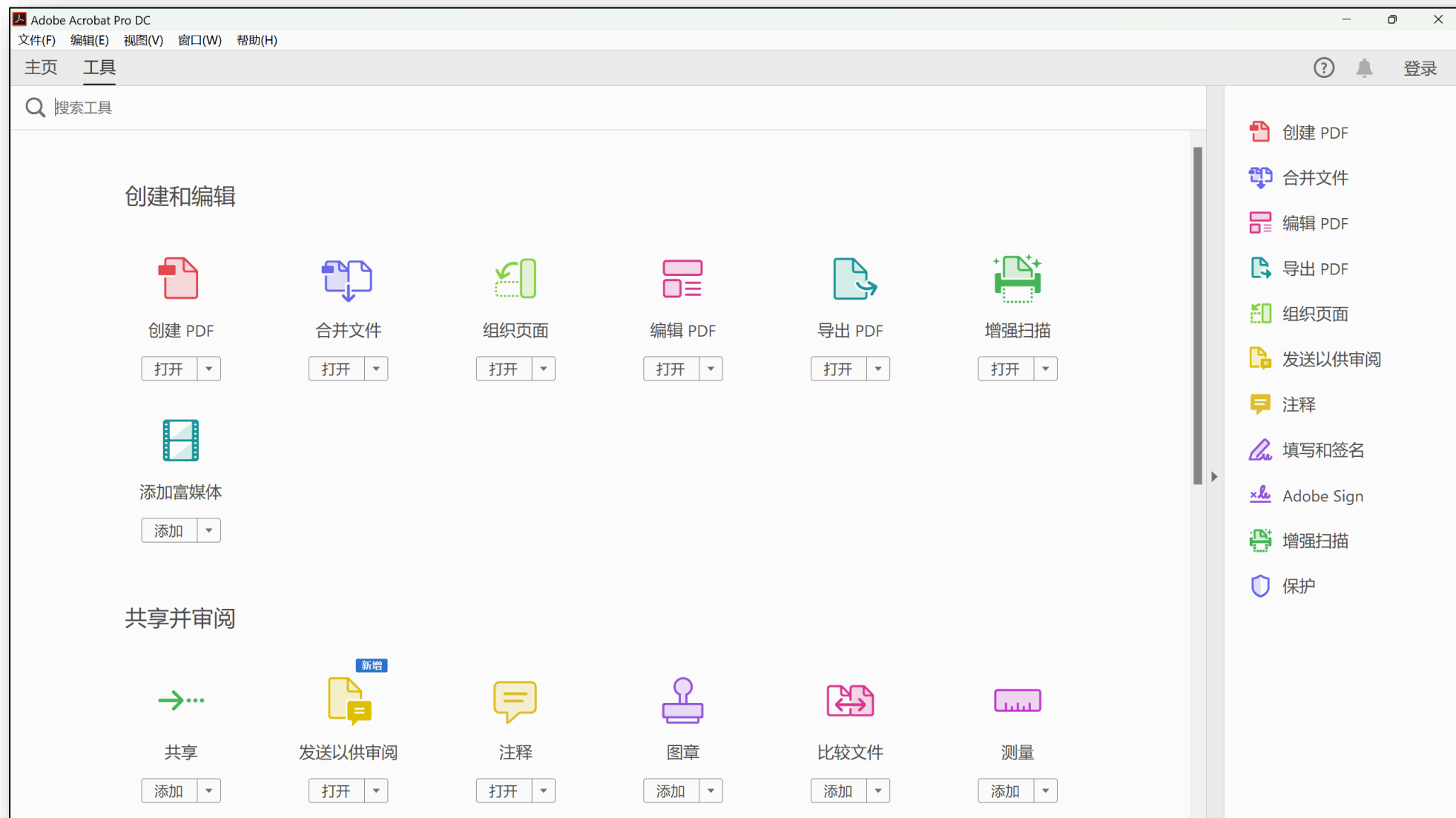
基本PDF功能:

- 合并+拆分
- 内容编辑
- 注释与签名
- 导出

Abbyy FineReader

精准OCR
+ 基本PDF功能

软件：Adobe Acrobat pro DC



案例1：文件合并为DPF

场景：你发表在杂志的论文需要提交给学校，要求提供杂志封面（JPG格式）、版权页（JPG格式）、目录页（JPG格式）、正文（PDF）、封底（JPG格式）。

要求：

按照顺序合并为一个PDF文件

案例2：PDF文档拆分与翻译

项目需求

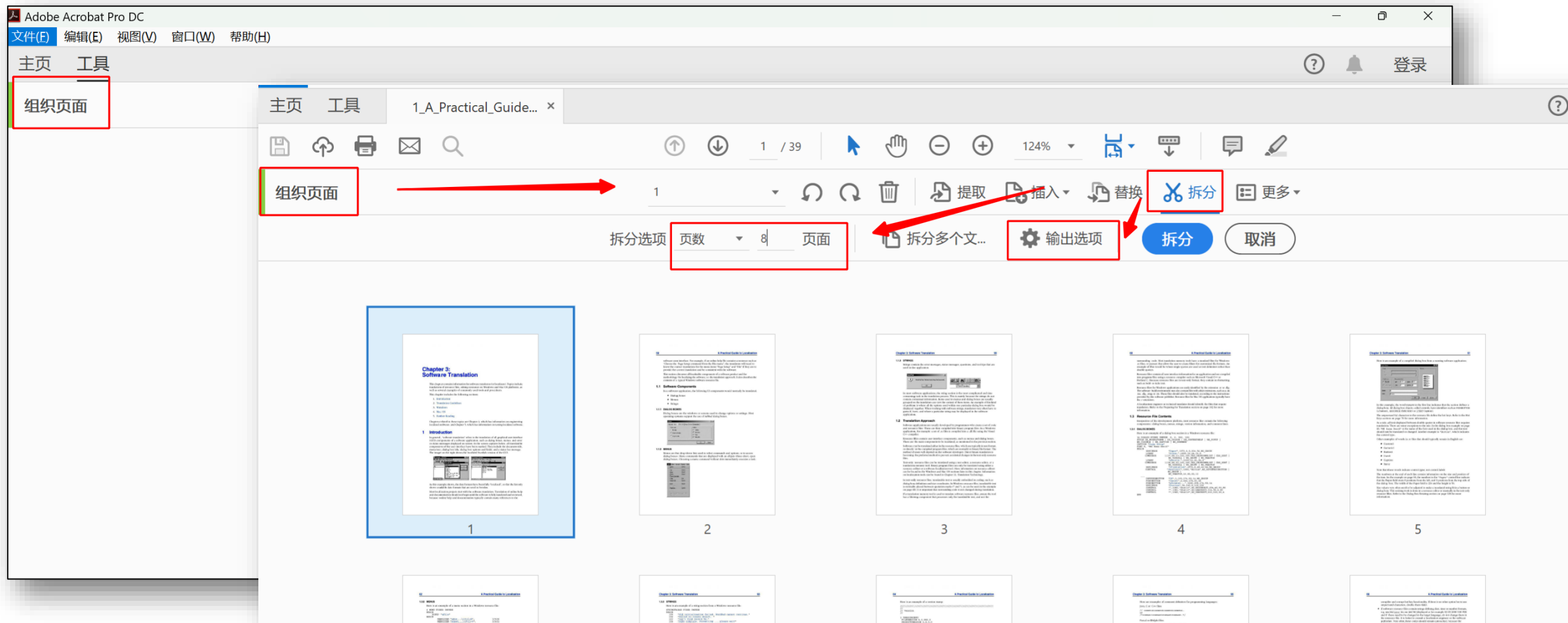
- 客户发来**39页**扫描的英文PDF文件，需要2天内翻译成PDF格式的中文文件。
- 当前翻译项目团队8人，其中5人做翻译，2人做校对，1人担任项目经理兼技术工程师。

项目挑战

- 如何将39页PDF拆分成5份？
- 如何将PDF转换成Word文件？
- 如何将翻译后的Word文件转换成PDF格式？

案例2: PDF文档拆分与翻译

Adobe Acrobat Pro DC



案例3：非图片格式文件的转换与翻译

目标：

- 将PDF文件转换为Docx文件

作用：

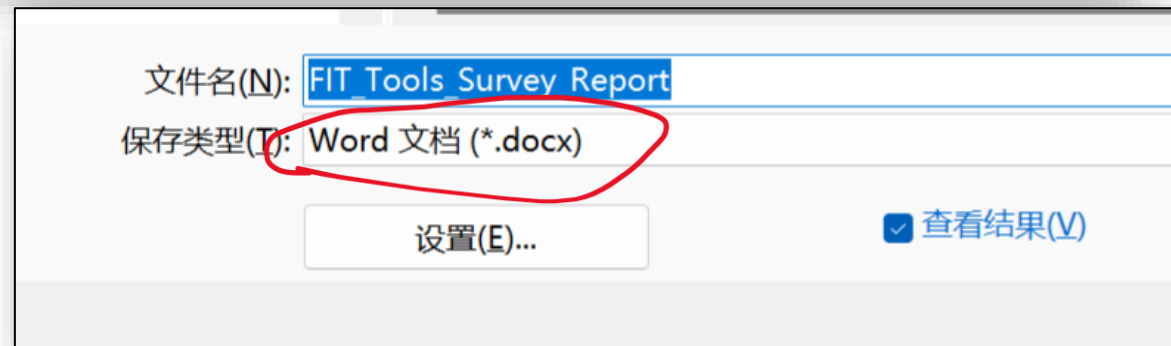
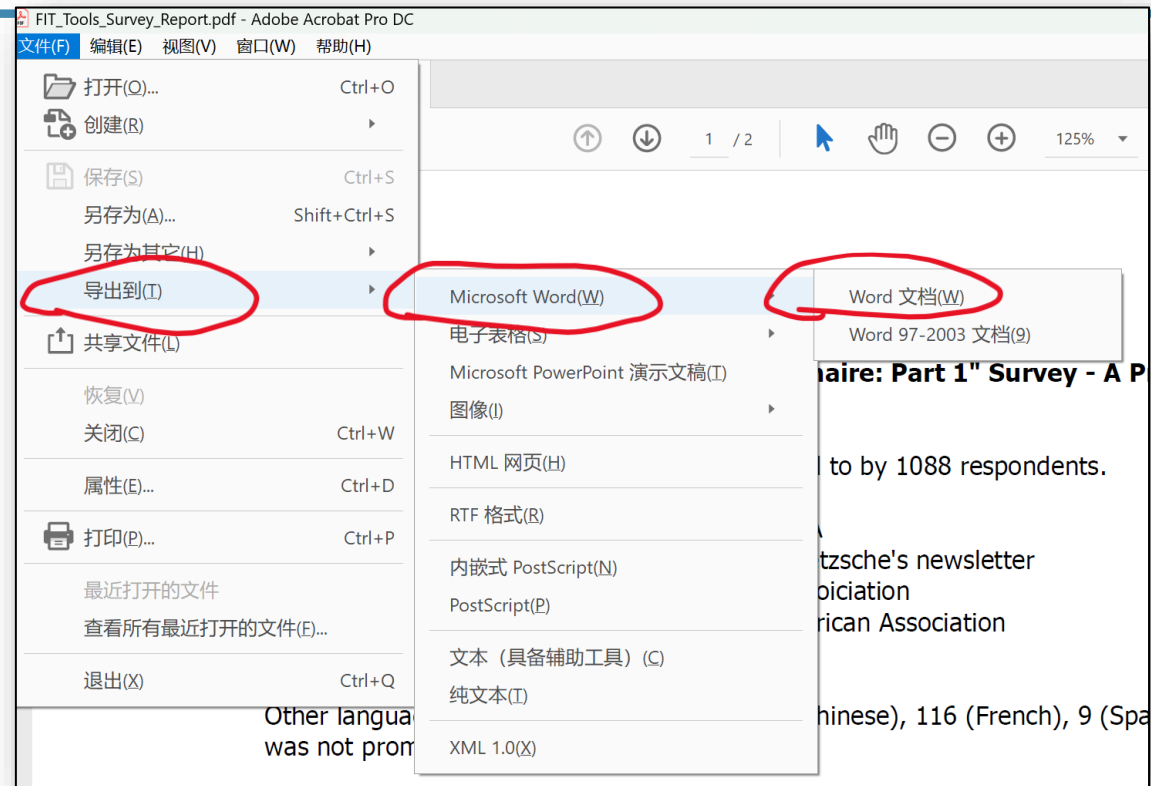
- 便于使用Word翻译

使用工具：

- Adobe Acrobat Pro DC

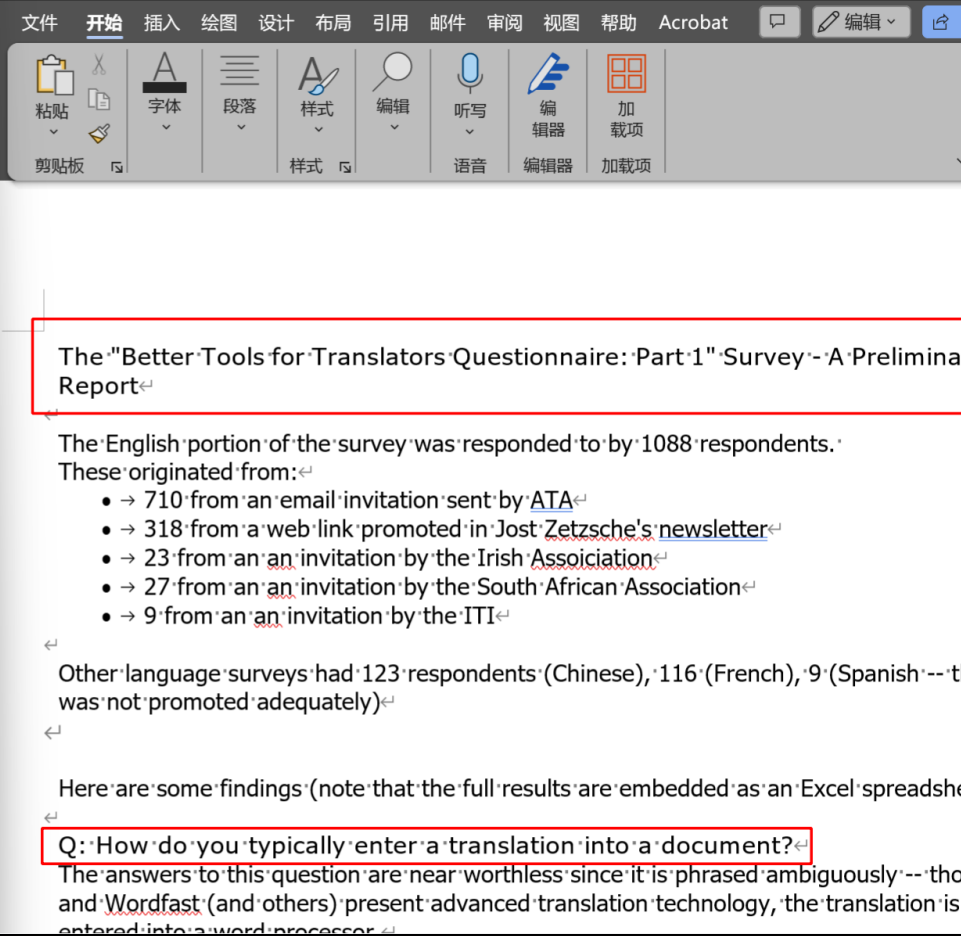
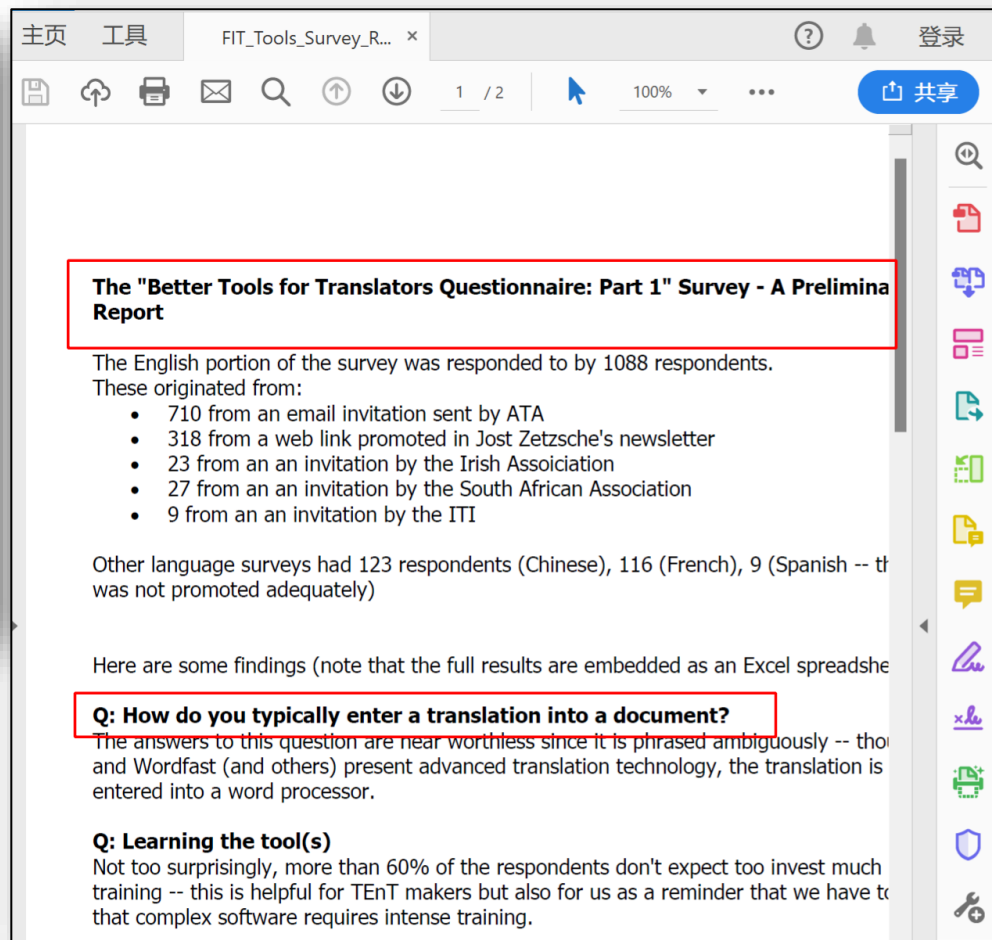
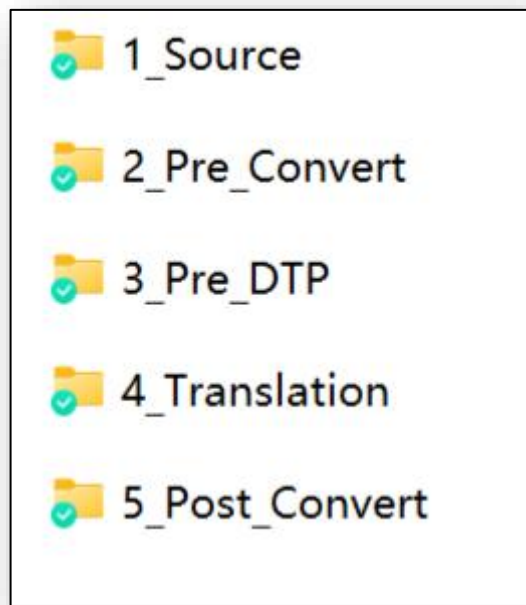
方法：

- 使用Adobe Acrobat Pro DC打开PDF，导出为word文档

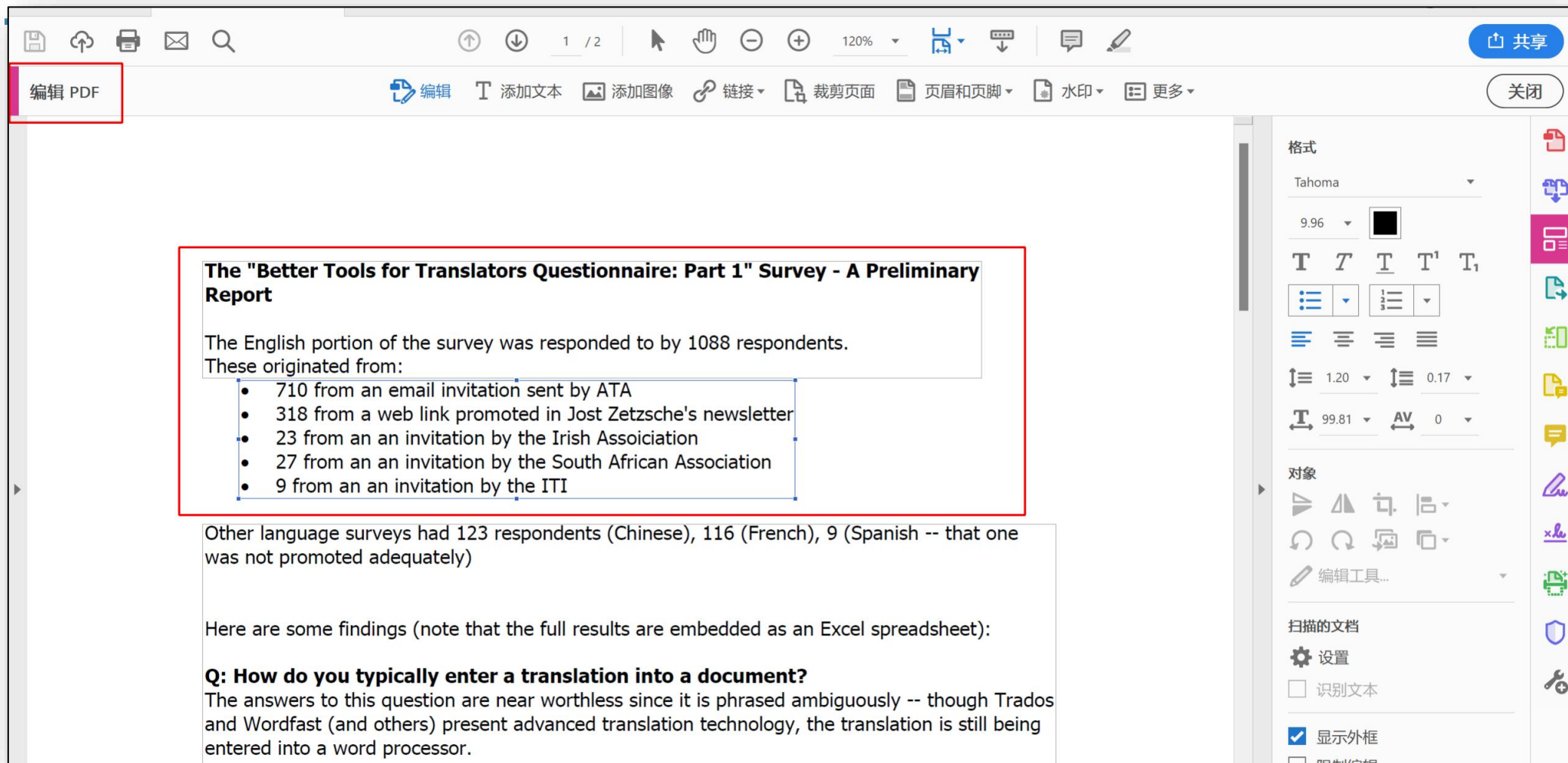


案例3：非图片格式文件的转换与翻译

DTP: Desktop Publishing 预排版



案例4：非图片格式PDF文件的编辑与修改



Adobe Acrobat DC 适用场景：少量文字编辑修改时

PDF文件的安全性设置

设置、删除密码

- 属性-安全性-安全性方法-口令安全性-要求打开文档的口令-文档打开口令

限制文档编辑和打印

- 设置允许打印的分辨率
- 设置可编辑更改的内容

加密所有文档内容

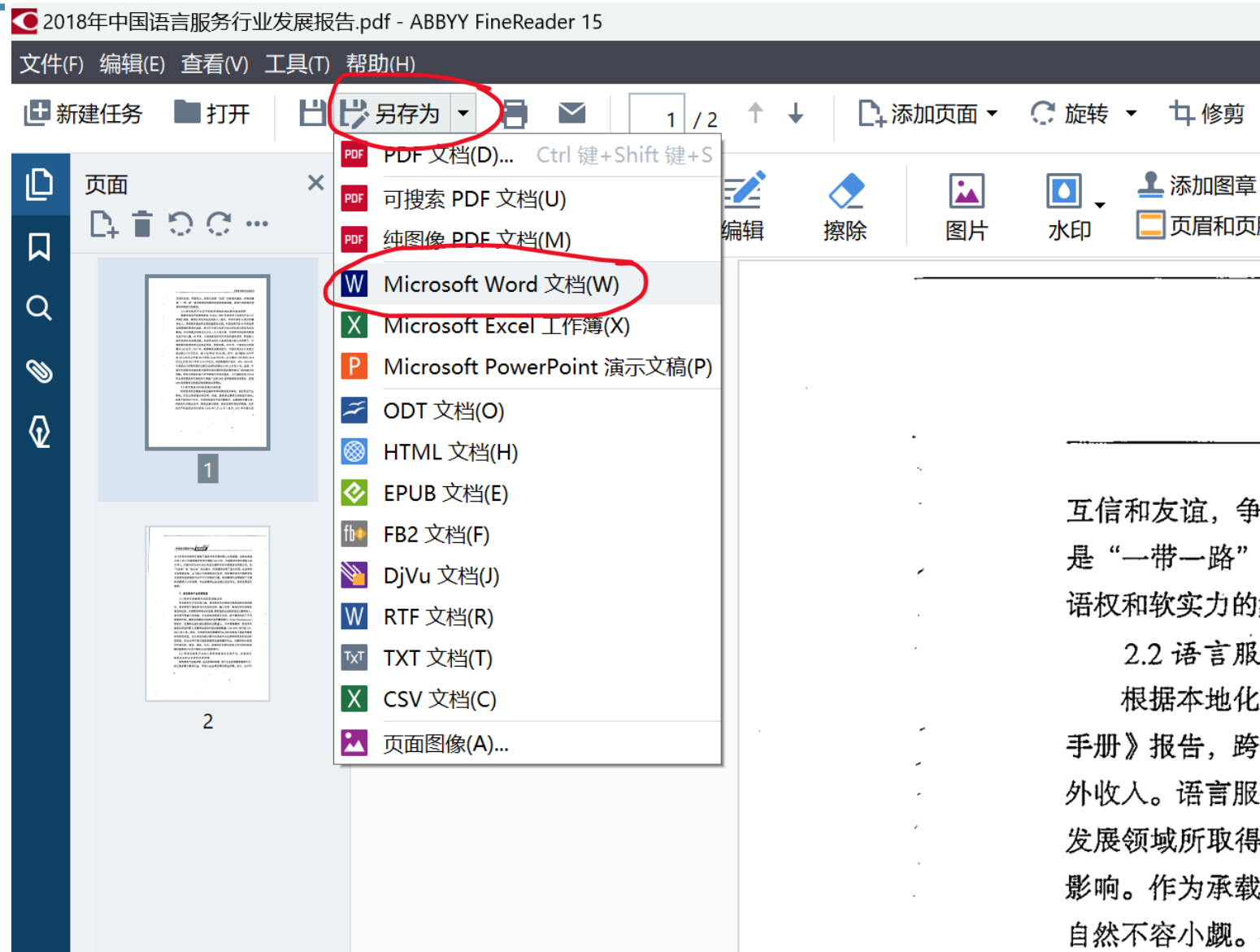
软件：Abbyy FineReader



案例5：图片格式的PDF文件转换为Word文件

用法1:

- Abbyy FineReader
- 进行全文档OCR
- 导出word文件



互信和友谊，争
是“一带一路”
语权和软实力的

2.2 语言服

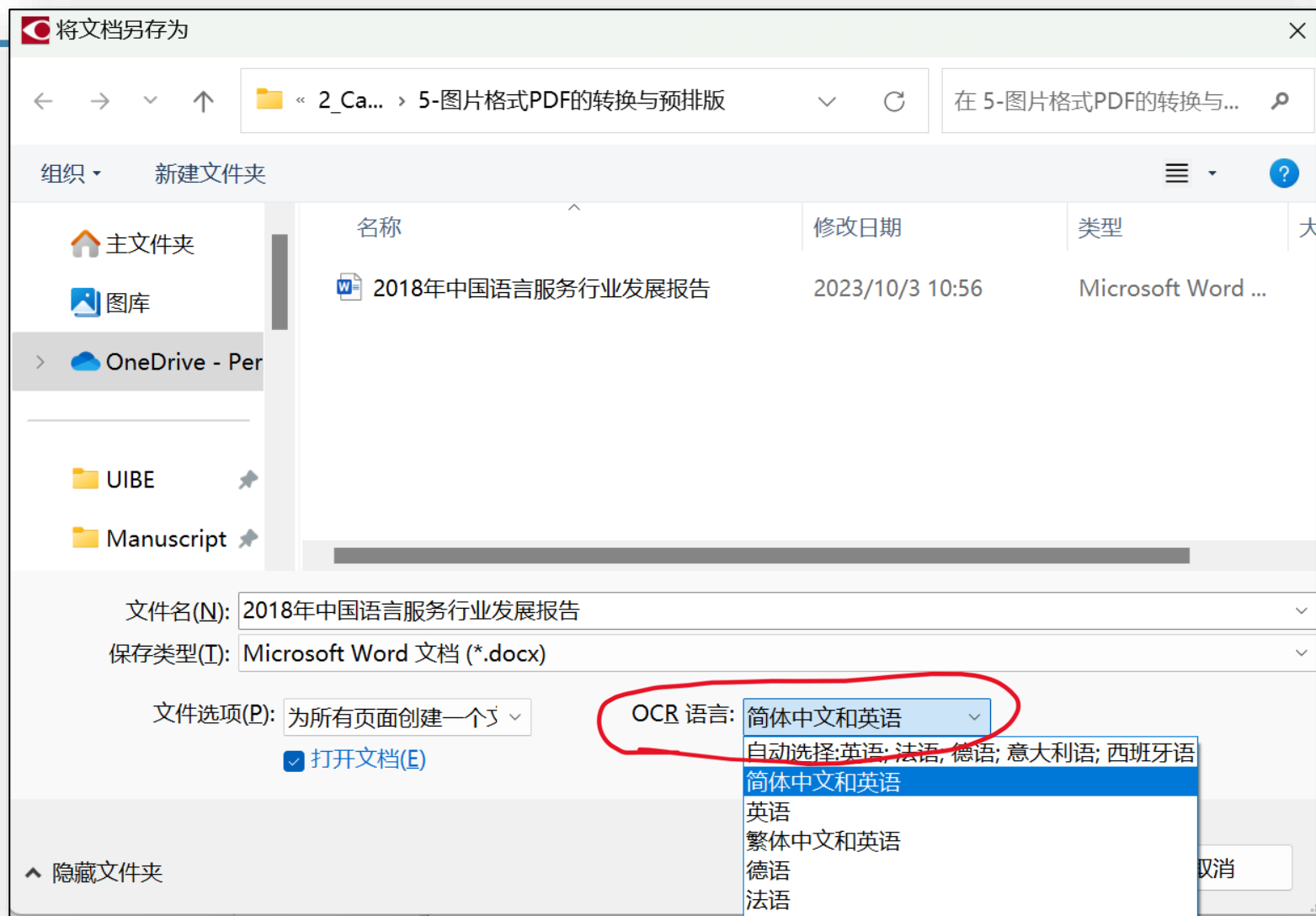
根据本地化

手册》报告，跨
外收入。语言服
发展领域所取得
影响。作为承载
自然不容小觑。

案例5：图片格式的PDF文件转换为Word文件

用法1:

- Abbyy FineReader
- 进行全文档OCR
- 导出word文件

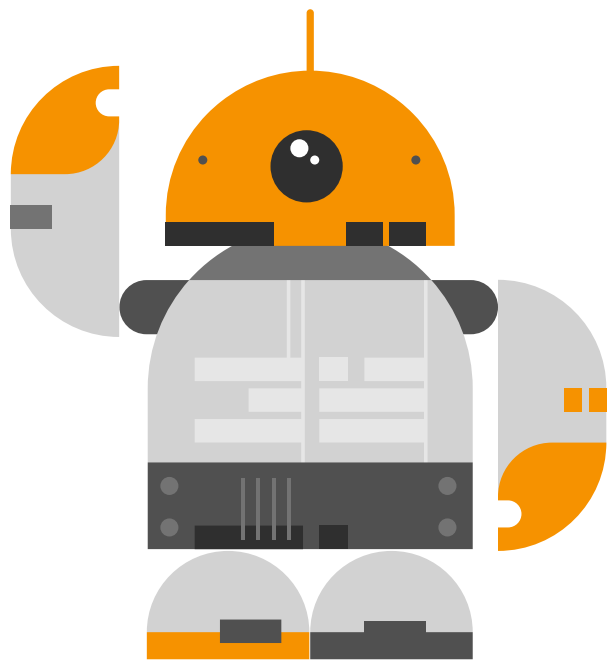


案例5：图片格式的PDF文件转换为Word文件



用法2:

- 使用 Abbyy Screenshot Reader 对屏幕区域截图
- 对少量文字进行OCR（字符识别）
- 结果保存在Windows剪贴板



5. 小组作业

小组作业

任务：

- 将图片格式PDF文件转换成DOCX文件，预排版，翻译和排版成PDF文件。

提交译文格式：

- 压缩作业文件夹Homework下的2-5文件夹为一个文件，不要修改文件名。
以CAT-小组名称-格式转换.zip命名

提交日期（Deadline）：

- 下次上课前1天

提交方式：

- 组长收集本组作业，每个组交一份，发给课代表，课代表发到
jie.huang@uibe.edu.cn（电子邮件字母不区分大小写）

END