目 录

第三章

双语语料库的建设与用途

本章先从双语语料库的概念及类型划分谈起，介绍双语语料库的建设步骤，包括建立原则、具体步骤、注意事项等；然后介绍国内外双语语料库的建设情况和双语语料库在翻译实践中的作用，包括双语语料库在语境关键词检索、翻译（术语）记忆和机器翻译方面的巨大作用；最后介绍基于语料库的翻译研究的情况，以及数字人文时代计算机工具在翻译领域的应用。

一、双语语料库的概念与类型划分

1961 年，世界上第一个机读语料库布朗语料库（Brown Corpus）在美国布朗大学诞生。从 20 世纪 80 年代开始，伴随着计算机技术的发展，出现了各种不同类型的第二代语料库，对语料库的应用开始上升到各个层面。双语语料库在第二语言习得、双语词典编纂、译员作品风格、机器翻译等领域中都得到了广泛的应用，成为与计算机辅助翻译联系最为紧密的语料库类型。双语语料库拥有大量原文与译文实例，能够在更大范围内方便译员或校对人员查找语言搭配、进行译文质量检查。双语语料库同译员培训结合起来则有助于译员通过语境探索，对全语境条件下的原文及译文进行考察，进一步加强对两种语言之间异同的理解。同时，也可以通过考察译员译文语料库评估译员培训的各个层面。[1]

Baker（1995）认为，双语语料库应分为三类。首先是平行语料库（parallel corpus），也称对应语料库，即通过将源语文本同译入语文本相对应后建成的文本库，对应层级包括单词级别、句子级别和段落级别。第二类是多语语料库（multilingual corpus），虽然被称作多语语料库，其实是由两个或多个不同语言的单语种语料文本通过同样的筛选规则搜集而成，不包括翻译文本，即不包括多个单语种的原文本语料库。第三类为可比语料库（comparable corpus），既包括源语文本，又包括从其他语言翻译为此种语言的文本。

1 Bowker（2003）曾详细论述了如何建立一个跟踪性语料库，以考察受训译员的翻译能力在某段时间内发生了多大程度的变化。

同另外两种双语语料库相比，与翻译实践（尤其是计算机辅助翻译实践）和教学联系最紧密的当属平行（对应）语料库。这种语料库中含有大量真实的翻译句子对，为翻译实践、翻译教学、翻译语言研究和语言对比研究提供了良好的基础。同时，平行语料库本身也是计算机辅助翻译术语库与记忆库的存在形式与载体。目前，国外已建成许多这样的双语平行语料库。早期著名的包括加拿大议会会议录英—法平行语料库（Canadian Hansard Corpus）、英语—挪威语平行语料库（English-Norwegian Parallel Corpus，ENPC），另外还有英语和意大利语、英语和德语对齐的平行语料库。目前，国内已建成的各种汉英平行语料库也属此类，如清华大学自然语言处理与社会人文计算实验室利用互联网平行网页获取软件和双语句子自动对齐软件得到的 THUMT 双语语料库，共包含 285 万汉英平行句对，属于句子级的平行语料库；绍兴文理学院建设的《红楼梦》汉英平行语料库，属于段落级的平行语料库；中国翻译研究院建设了政治类文献的汉英双语语料库，[1] 也属于段落级的平行语料库。

在平行语料库下，还可以进一步细分为单向对应语料库（unidirectional parallel corpus）、双向对应语料库（bidirectional parallel corpus）和多向对应语料库（multidirectional parallel corpus）。

单向对应语料库指整个语料库都是由一种语言翻译到另外一种语言所构成的对应语料库。在计算机辅助翻译中，单向对应语料库是翻译记忆中最普遍的形式，即将已经翻译过的文本同原文对应，然后做成对应语料库为将来的翻译任务服务，两种语言之间不能相互自动转换。

双向对应语料库指整个语料库中包括原文和译文文本，如著名的 Canadian Hansard Corpus（英语—法语平行语料库），ENPC（英语—挪威语平行语料库）。北京外国语大学的中国英汉平行语料库也是双向对应平行语料库，在语料库搜索栏中输入中英文单词后，出现的是对应的汉译英和英译汉的两种句子对，如输入“穷尽”，得到的结果既包括中文里含有“穷尽”和相应的英文句子对，也包括原文是英文，其中文译文含有“穷尽”的句子对。当然，可以选择是否只需要一种结果。图 3.1（Facchinetti 2007: 53）展示了 ENPC 这个典型的双向对应语料库的结构，是由 1993 年的 ICAME 会议上展示的模型修正而来，其中双箭头表示可以进行双向研究。

1 检索地址为 http://www.china.org.cn/chinese/catl/node_7232138.htm，2023 年 7 月 24 日。

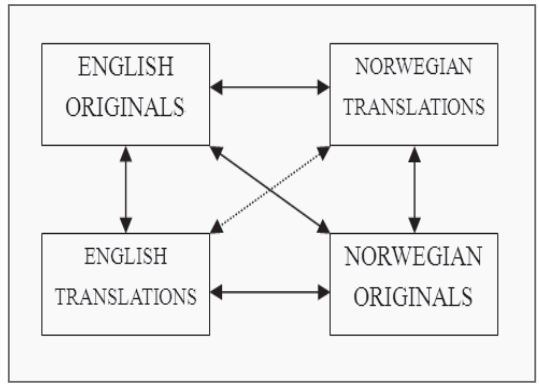


图 3.1　ENPC 双向对应语料库结构

多向对应语料库指整个语料库是由同一种原文的文本及其他语言的译本组成。图 3.2（Facchinetti 2007: 54）为多向对应语料库建设的钻石结构，虽然设想甚好，然而这样在三语之间的译文文本极其有限，文本类型也不容易匹配。

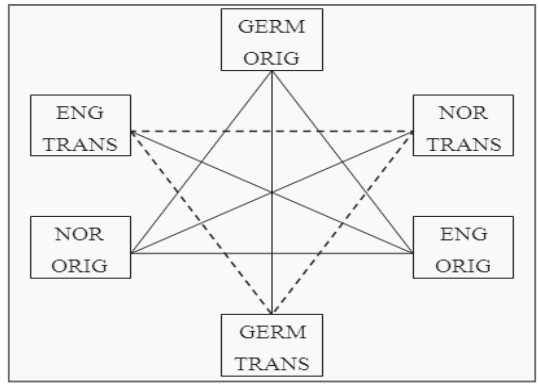


图 3.2　多向对应语料库结构

对于平行语料库，其他学者也有不同的划分方式，如表 3.1（Tognini 2001: 7）。

表 3.1　平行语料库划分

Parallel Corpus	Translation Relationship	Alignable
Free-Translation Corpus	Translation Relationship	Not Alignable
Comparable Corpus	No Translation	Not Alignable

二、双语语料库的建设

Sinclair（1963）曾指出，“任何语料库研究都必须以建立一个相应的语料库为前提，决定语料来源和语料搜集过程几乎为之后所有的研究工作奠定了基础”。对翻译而言，建立双语语料库对翻译的各个方面都有非常重要的意义和用途。语料库的建立一般都会涉及两个问题，即代表性问题和文本取样策略问题（包括随机取样或分层取样、采取内部标准或外部标准、采集全部文本或部分文本等）。对于 CAT 软件来说，建设双语平行语料库就是为翻译任务作准备。因此，搜集双语语料的工作一般从两个角度进行：一是构建术语库，二是构建翻译记忆库。简单说来，前者注重词汇，后者注重句式。术语库的建设通常不涉及对齐或标注等问题，可以直接从相关领域中汲取专有名词或固定说法，亦可在翻译过程中定义术语并添加到已建成的术语库当中。对于构建翻译记忆库而言，最重要的就是对齐（alignment）和双语检索（bilingual concordancing）。因为前者同翻译记忆密不可分，后者对译文选择至关重要，两者都将对翻译实践产生不可忽视的辅助作用。[1]

双语语料库的建设通常要先确定建库目的，然后搜集双语材料，并进行进一步的加工处理，再对语料进行对齐后入库。为了充分利用语料库，必须开发专用检索软件，后期的库维护也必不可少。

1．按需确定建库目的

建库目的决定了双语语料库建立的整个过程。所以，明确目的对后续工作将会产生非常重大的影响。Johansson（1991: 305-306）曾指出，虽然大型语料库具有无可置疑的优势，然而构建能从多角度透彻分析的小型语料库也有其意义。国家级的双语语料库一般都力求完备，尽量搜集各个领域内的可靠译文，但会根据其研究或应用目的确定语料的时间、语域，尤其是文本类型范围。翻译公司的业务范围可能集中于某几个领域，如法律、经贸、机电、医学等，所需的双语语料库类型一般都集中在某个领域内。而一般的个人译员会对某一个自己擅长的翻译领域感兴趣，不太可能精通所有领域。例如，在经贸范围内就存在进出口业务、银行业务、股票市场等更细致的划分。对于译员来说，可以根据自己的实际需要，明确建库目的，不单纯追求语料范围和大小，而是尽量搜集某领域内的可靠译文。

1　对齐与翻译记忆详见本书后面有关章节。

2．搜集双语材料

搜集语料必然涉及语料的代表性问题，即所搜集的语料在所选定的研究范围内是否具有可靠的代表性。语料库的设计和单独文本的选择，都取决于建立语料库本身的目的。对于纯粹的翻译研究来说，双语语料库的建设可能是要研究某种特殊句式（如被动句的翻译），因而在搜集语料的过程中就要搜集大量的相关材料，并注意某种体裁或者句型的整合。

在译员接触新的翻译工作之前，需要找相关领域的双语文本材料进行训练，以便熟悉相关术语和句式的翻译。句式固定的文本材料最适合作为翻译记忆使用，因为当新的翻译任务出现时，可以通过翻译记忆产生译文。在翻译实践中，译员经常会遇到令人手足无措的专业术语，即在一定的领域内有特殊含义而且形式固定的语言表达。训练翻译软件对这些固定的术语、表达方式和句式进行识别，即做好双语语料库，译员就可以不用花费大量时间去熟记相关领域的术语，而是将更多的精力放在提高译文质量上，进而提高实际翻译中的效率。

对于不同的领域，搜集双语材料的方法可能不同，如对于学术论文摘要翻译，可能从单纯的中英文论文对应摘要中搜集材料后对齐，还不能满足高质量的译文需求。如果要从事某领域的科技论文翻译，则需要对这个领域有所了解，同相关专家讨论后做出可靠的译文后再对齐，这样建成的双语语料库才有可能真正应用于最终的翻译实践。

3．语料处理

通常，对齐后的双语语料库可以直接为计算机辅助翻译服务。语料处理与翻译实践联系不大，但如果要利用双语语料库检索进行译员培训，则对语料进行预处理就是前期建库工作中必需的步骤。根据语料库应用目的的不同，处理语料有很多种方法，与实践联系最为紧密的是对汉语的切分。按照汉英双语语料库的建设和研究目的，还可能涉及对词性、句子结构、文本结构及文本来源等各项进行标注。本节只介绍与译员检索联系紧密的汉语切分处理。

汉语本身是一种缺乏单词形态变化的语言，词的类别不能像印欧语系语言那样能直接从词的形态辨别（刘开瑛 2000：3）。所以，汉语的分词不能像英语等印欧语系语言那样，直接通过单词之间的空格辨别。对应语料库的建设离不开汉语语言理解，而汉语语言理解又离不开对输入文本进行句法分析（parse）。在计算机辅助翻译译员培训当中，需要使用到双语语料库进行双语检索，只有进行分词之后，检索才能成功进行。一般认为，计算机从事句法分

析所凭借的语法知识只能来自句法规则库，而这些规则一般都是建立在词法和语义知识之上的。因此，必须先对汉语句子进行词汇切分处理后，才有可能进行句法分析。汉语中“词”的概念很模糊，随着建设大规模、高等级汉语语料库的呼声渐高，汉语书面语的分词技术已经形成了一门具有挑战性的新兴学问。目前，为了满足信息处理的需要，我国已出台《信息处理用现代汉语分词规范》，详细规定了现代汉语的分词原则，对汉语信息处理的规范化和各种汉语信息处理系统之间的兼容性有着重要的作用。

为了更加清楚地解释自动分词的概念，这里选取叠词较多的朱自清散文名作《春》作为分析对象，用 ChatGPT-3.5 Turbo 版本进行处理演示。图 3.3 展示的是《春》前三段的处理结果，整句的中文被切分为不同词语与句段的组合。中文下方的部分标注及含义为：VV（动词）、PU（标点符号）、NN（名词）、VA（谓语形容词）、CD（基数词）。分词后才可进行词汇与搭配检索。

```
盼望 着 ， 盼望 着 ， 东风 来 了 ， 春天 的 脚步 近 了 。
VV AS PU VV AS PU NN VV AS NN DEG NN VA SP。
一切 都 像 刚 睡醒 的 样子 ， 欣欣然 张开 了 眼 。
CD AD VV AD VV VV DEC NN DEG NN VA AS NN SP。
山 朗润 起来 了 ， 水 涨 起来 了 ， 太阳 的 脸红 起来 了 。
NN VA VA AS SP NN VV VA AS SP PU NN DEG NN VA AS SP。
小草 偷偷地 从 土里 钻 出来 ， 嫩嫩 的 ， 绿绿 的 。
NN VA AD P NN VV VV VV PU VA SP PU VA SP。
园子 里 ， 田野 里 ， 瞧 去 ， 一大片 一大片 满 是 的 。
NN LC PU NN LC PU VV VV VV PU CD CD VA DEC SP。
坐 着 ， 躺 着 ， 打 两个 滚 ， 踢 几 脚球 ， 赛 几 趟 跑 ， 捉 几回 迷藏 。
VV AS PU VV AS PU VV CD M NN PU VV CD M NN PU VV CD M NN PU VV CD M VV PU VV CD M VV PU VV CD M VV SP。
风 轻悄悄 的 ， 草 绵软软 的 。
NN VA VA DEG PU NN VA VA DEG SP。
```

图 3.3　分词效果展示

4．双语语料库对齐

在双语语料库建设中，要考察相关原文在具体语境中的译文，最重要的一步就是对双语文本进行对齐。对齐主要是对双语材料句子级别的对齐，但为了让记忆库在计算机辅助翻译中起到更大的作用，一般都会将材料划分为比句子单位更小的级别。专门的双语对齐术语库也从另一方面体现出双语语料库在翻译实践中的巨大作用。面对大量的对译单位，人工完成对齐虽然精确率高但是效率低，难以建成大规模的双语语料库。因此，通常使用软件自动对齐局部对译单位，然后采用人工调整的方法检查是否对齐。

在过去几十年中，众多的研究者在双语语料库对齐方面进行了大量的研究。对齐方法基本可以分成基于句子长度的方法（Brown *et al.* 1991; Gale & Church 1991）、基于词汇对应的方法（Kay & Roscheisen 1993）和混合法（Tan & Nagao 1995）。基于句子长度的方法指利用原文句子与其译文句子在长度上存在的一定关联进行对齐。在实现的过程中，通常是先人工建立一个双语语料库作为参照，计算原文和译文中句子长度的比例，然后通过选取先建立好的双语库中的匹配概率，确定其对应类型（如 1:1 或 1:1.5）。另外，Kay & Roscheisen（1993）还提出了基于词汇信息的模型，即在两个对应文本中搜寻对译的单词，如果一对句子出现足够多的互译单词配对，那么整个句子就判断为对译句子，可以对齐。简单说来，最佳的句子对是那些使系统词汇对齐数量最大化的句子。这样，通过对齐几类词，如代词、数词和专有名称等，就会在两个文本中找出相对稳定的对译关系。国内关于双语对齐的研究提出了基于词汇之间的相互关联度，进行多次组合的识别方法，并利用先假设再检验的方法在双语语料库中抽取翻译等值单位（常宝宝 2002）。

Chuang 等（2005）提出了基于标点符号的对齐方法，并对 Chinese-English Sinorama Magazine Corpus 平行语料库进行了对齐，结果发现基于标点符号的方法比基于句子长度的方法更加准确有效，其准确率超过了 93%。这一发现在实践中得到了广泛应用。目前，主流的计算机辅助翻译软件提供的对齐工具（如 Trados 中的 WinAlign 和 DéjàVu X 中的 Alignment Workfile 等）都是依靠标点符号将两边的文本分割成许多对译单位，允许译员将两个对译的文本导入对齐，人工修正后导入翻译记忆。由于汉英语言差异，除了特定范围内的文本（如政治、科技、法律、贸易等）易于对齐（即由软件对齐后不需要人工再进行大量修改工作），大部分文本在软件协助对齐后，还需要人工进行费时费力的修正。由 Michael Barlow 研发的 ParaConc 是一款多用途的语料库软件，它允许对齐后进行检索、查找、翻译、多语支持、搭配频率统计、高级（分类）检索等，支持通配符检索和纯文本格式的文件，可以灵活定义语言和检索行的大小，实现双语平行语料库检索，其详细使用步骤参见本书第四章。以中国银行 2007 年年报双语目录为例，经过软件处理后可以得到如图 3.4 的对齐文件。

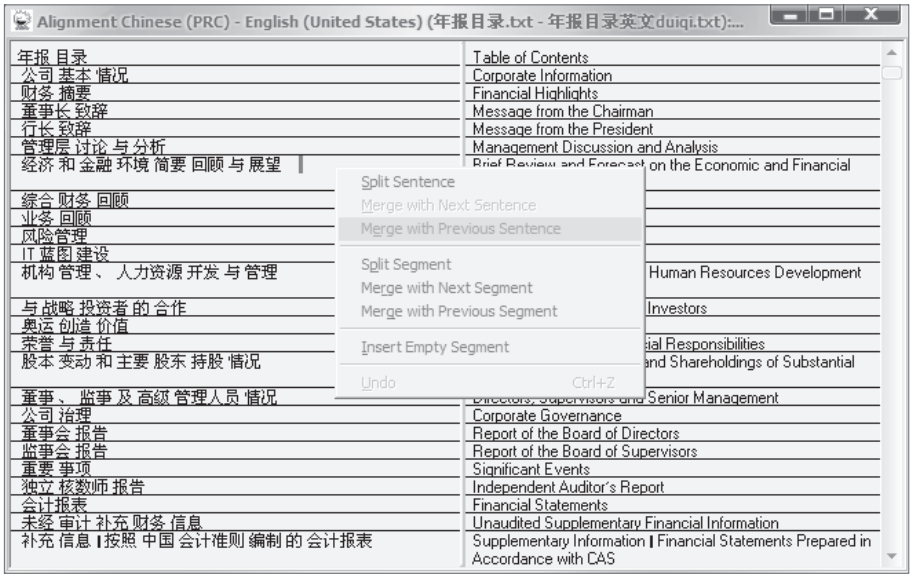


图 3.4 对齐效果展示

需要注意的是，由于两个目录文本已经提前经过分行处理，所以对齐结果无需人工修正。在实际情况中，自动对齐不可能达到绝对匹配的效果，需要通过图中选项卡所列出的选项进行修改，如 Merge with Previous Sentence（同上一句合并）等。实际上，双语文本的对齐工作要复杂得多。语言翻译是相对复杂的过程，任何一种单一的概率统计方法都不可能圆满处理复杂的情况，所以到目前为止，任何单纯依赖软件的双语对齐都没能在大范围内取得令人满意的效果。

现在，基于大量的语料库资源及相关工具，用户利用语料库工具对所搜集语料进行双语（或多语）对齐的结果一般已经能达到 75%—95% 的准确度。在计算机辅助翻译研究与实践中，学界普遍认为，达到 50% 或以上句子层面对齐的双语（或多语）语料才有使用价值。用户可以对计算机辅助工具里的对齐选项进行设定，将其内置的对齐（匹配）率设置为至少 50%。设置的匹配率值越高，所检索出来句子的可利用率越高，对译者的辅助作用也就越大。

5．检索软件开发与后期库维护

库维护的概念其实贯穿了语料库建设的整个过程。在完成了双语语料的初步自动对齐之后，需要对所完成的工作进行人工检查和确认。同时，必须开发相应的检索软件以便研究使用。小型的语料库可以利用现有的语料库检索软

件，如前文所提到的 AntConc 等。然而，大型的语料库一般含有多方面的研究目的，这些软件往往不能满足研究需要。所以，大型的双语语料库一般都会开发忠于自己建库目的的语料库检索软件。例如，TEC 语料库的客户端检索程序就是由都柏林三一大学的 Saturnino Luz 博士专门设计的。本章第四节介绍双语语料库应用时会详细讲解基于 TEC 的翻译研究。

技术维护同后期库维护相辅相成。对于研究者而言，某一研究范畴内的语料搜集务必力求完备。因此，双语语料库的建设不是一项一蹴而就的工作，而需要不断丰富语料并进行多方位标注。所以，如果想让语料库一直保持有效的利用价值，就必须重视后期的维护工作。语料库的成功建设和维护同充足的资金支持密不可分。语料库建设的要事之一就是得找到充足的运转资金（Baker 1999: 284），否则，长期的建设和研究工作不可能稳定地进行下去。

三、双语语料库的应用

随着语料库语言学研究的深入发展，双语语料库正在得到越来越广泛的关注与应用。对语料库和翻译的研究一般集中在理论与实践两方面（Hunston 2002: 123）。从理论上来讲，语料库为翻译过程的研究提供了原材料。从实践上来说，研究主要聚集在如何开发一种让译员能利用双语语料库来进行翻译的软件。换句话说，双语语料库最重要的用途之一是译员可以看到对齐好的前人所做的译文，同时也让翻译研究者有机会考察在单语语料库中难以发现的语言之间的对应和差异。

在同翻译相关的活动和研究中，平行语料库有多大作为呢？ Tognini（2001）指出，平行语料库的最大好处就是将翻译好的成品展现给人们，从这些成品中人们可以了解不同语言之间的异同。朴松林认为，基于双语语料库的应用领域主要包括语料库语言学、双语词库的提取、词典编纂和语言工程，如机器翻译等（王克非 2004）。双语语料库中存储的大量真实翻译实例不仅为译员培训提供了良好的素材，译员本身产生的译文也可以集中成语料库，用以对译员进行考察。

语料库，尤其是双语平行（对应）语料库，可以为机器翻译提供真实的译文实例，从而帮助改进机器翻译系统。同时，双语语料库中的对应文本还有助于进一步认识翻译过程。对于翻译实践来说，语料库还可以作为翻译人员的参考工具，帮助译员提高翻译质量。总的来说，译员培养、双语词典编纂及机器翻译是双语语料库最普遍而广泛的应用范畴。

1．国内外双语语料库检索

大型的双语语料库可以从词语搭配、术语规范等方面进行多方位的译员培训。下面介绍几个著名的双语语料库。

（1）北京外国语大学的 CQPweb 语料库[1]

利用语料库进行翻译研究在我国起步较晚，但是发展很快，目前国内最热门的大型语料库包括北京外国语大学的 CQPweb 语料库。此语料库搜集量大，包括小说等各种题材，并且免费提供了测试版本。用户可以选择其中的双语语料库，此处以选择 Yiyan 语料库为例，其检索界面如图 3.5 所示。双语平行语料库以其大量的对译材料为语言和翻译研究提供了新的途径，图 3.6、图 3.7 分别展示了检索项为“新时代”和 party 的检索结果。

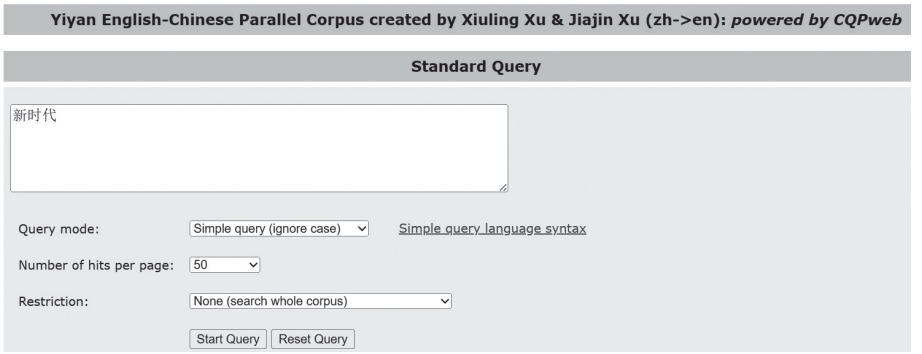
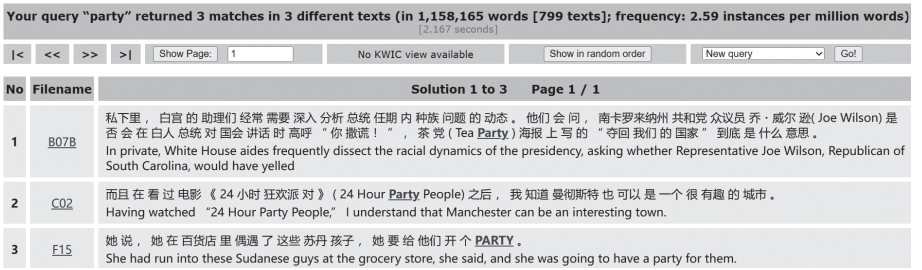


图 3.5　Yiyan 检索主页面

5	F39B	我 知道 你 在 想 什么 ： 这 听 起来 像是 某种 奇怪 的 **新时代** 理念 。 I know what you're thinking: That sounds like some bizarre new-age philosophy.
6	G53B	“ 不同 的 数据 挖掘 模式 可以 把 人类 带领 到 一个 科学 发现 更加 迅速 的 **新时代** 。
7	G56	罗马 吉普赛人 ， 大约 1000 年前 发源于 印度 次 大陆 ， 现在 遍及 欧洲 ； 爱尔兰 流浪者 ， 他们 有 共同 的 语言 ， Shelta ， 他们 被 认为 是 在 16 或 17 世纪 成为 游牧 民族 的 ； 再 加上 **新时代** 的 流浪者 ， 嘻皮士 和 地壳 朋克 （ crusties ） 。 Roma Gypsies, who originated from the Indian subcontinent around 1,000 years ago and have now spread across Europe; Irish Travellers, who have a common language (Shelta) and are believed to have became nomadic in the 16th or 17th century; plus new age travellers, hippies and crusties.
8	G59A	在 亚利桑那州 ， 他 与 不 受 人 喜欢 的 环境 作家 和 行动主义者 爱德华·艾比 成为 朋友 ， 并 凭借 个人 能力 最终 成为 一个 有名 的 人物 。 用 《 凤凰 **新时代** 》 周报 的话 说 他 是 一个 身高 6.4 英尺 （ 约 1.95 米 ） 的 “ 尼尔·杨 和 罗伯特·米彻姆 的 变体 ” 。 In Arizona, he befriended the gadfly environmental author and activist Edward Abbey and became a well-known figure in his own right, a 6-foot-4
9	J36_e	它 无处不在 各国 人民 心中 产生 一 种 隐隐约约 的 **新时代** 即将 来临 的 概念 ， 一 种 变革 与 改良 的 朦胧 希望 ； 但 谁 也 猜 不 出 大革命 究竟 应该 是 什么 样子 。 It made people everywhere think that new times were coming and stirred vague hopes of change and reform, but no one yet suspected what it was to become.

图 3.6 “新时代”的检索结果

1　检索地址为 http://114.251.154.212/cqp，2023 年 7 月 24 日。



Your query "party" returned 3 matches in 3 different texts (in 1,158,165 words [799 texts]; frequency: 2.59 instances per million words)
[2.167 seconds]

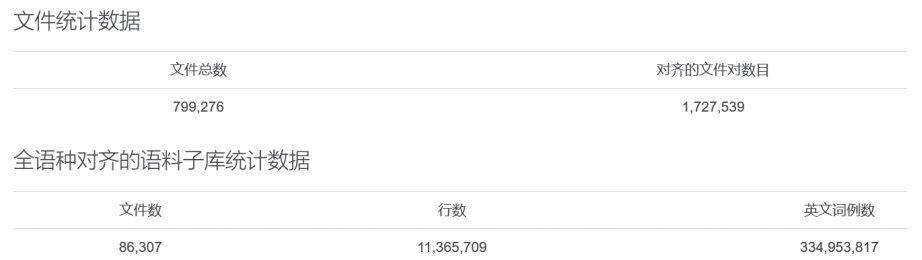
|< << >> >| Show Page: 1 No KWIC view available Show in random order New query Go!

No	Filename	Solution 1 to 3　Page 1 / 1
1	B07B	私下里， 白宫 的 助理们 经常 需要 深入 分析 总统 任期 内 种族 问题 的 动态 。 他们 会 问， 南卡罗来纳州 共和党 众议员 乔 · 威尔 逊(Joe Wilson) 是 否 会 在 白人 总统 对 国会 讲话 时 高呼 " 你 撒谎 ！ " ， 茶 党 (Tea Party) 海报 上 写 的 " 夺回 我们 的 国家 " 到底 是 什么 意思 。 In private, White House aides frequently dissect the racial dynamics of the presidency, asking whether Representative Joe Wilson, Republican of South Carolina, would have yelled
2	C02	而且 在 看 过 电影 《 24 小时 狂欢派 对 》(24 Hour Party People) 之后 ， 我 知道 曼彻斯特 也 可以 是 一个 很 有趣 的 城市 。 Having watched "24 Hour Party People," I understand that Manchester can be an interesting town.
3	F15	她 说 ， 她 在 百货店 里 偶遇 了 这些 苏丹 孩子 ， 她 要 给 他们 开 个 PARTY 。 She had run into these Sudanese guys at the grocery store, she said, and she was going to have a party for them.

图 3.7　party 的检索结果

（2）联合国平行语料库 [1]

该语料库在 2023 年提供的版本包含 1990 至 2014 年编写并经人工翻译的文字内容，包括以语句为单位对齐的文本。其双语对齐的文件涵盖阿拉伯语、西班牙语、法语、俄语、汉语和英语共 6 种联合国官方语言。具体数据如图 3.8。



文件统计数据

文件总数	对齐的文件对数目
799,276	1,727,539

全语种对齐的语料子库统计数据

文件数	行数	英文词例数
86,307	11,365,709	334,953,817

图 3.8　联合国平行语料库数据（2023 年）

2．平行语料库与翻译教学：语境关键词检索

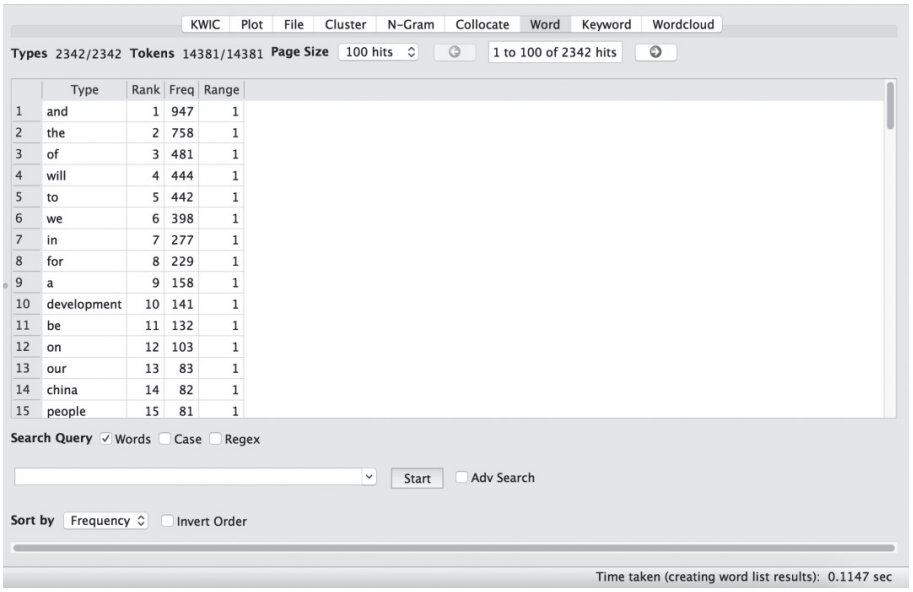
显而易见，人们在翻译时不会脱离语境来逐字翻译，而是将词置于足够大以至于没有歧义的意义单位（unit of meaning）中来考虑译文，即将几个词的组合作为一个翻译单位来翻译。双语语料库指导翻译实践的体现之一是在译员培训或翻译实践中，让译员查找语境关键词（Key Word in Context，KWIC）。大部分翻译单位都是由意义模棱两可的单词及其语境所构成，而有歧义的单词可以借助所在的语境使意义变得清晰。所以，在翻译实践和译员培训中，译员经常需要检索某个词在语境中的具体用法，以此来确定自己的译文是否地道合

1　检索地址为 https://conferences.unite.un.org/uncorpus/Home/Index/zh#statistics，2023 年 7 月 24 日。

法，在专业领域如法律、技术、医学等更是如此。而基于双语语料库的翻译对等研究，正好为译员提供了相应的培训材料。

在翻译教学方面，双语语料库的作用主要有（王克非 2004）：为某一检索词或短语提供丰富多彩的双语对译样本；为常用结构提供多种双语的对译样例，便于讲授者讲解及学习者模仿；提供丰富的可随机提取的一部分多译资料作为对照参考。另外，针对某一内容、某些专题和特定领域，还可以对译员进行翻译策略的培养。当然，在这些特殊领域中，搜集资料和编辑术语表时，双语语料库十分便捷有效。

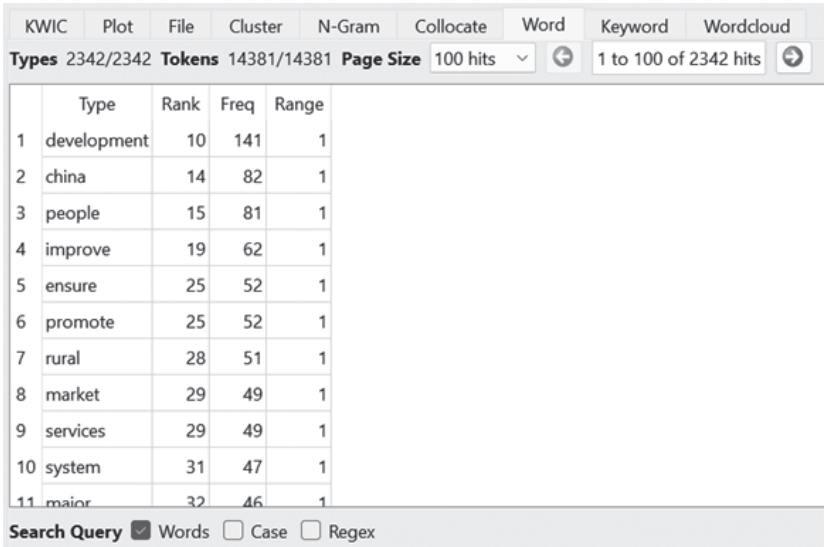
下面以 2022 年中国《政府工作报告》的英文版为例进行分析，所使用的软件为 AntConc 3.2.1。首先打开软件，选择主菜单 File 下的 Open File 选项，在导入想要分析的文本文件（2022 年中国《政府工作报告》）之后，选择 Word List 选项卡，便可以得到相应文本文件的关键词词频统计表，如图 3.9。



	Type	Rank	Freq	Range
1	and	1	947	1
2	the	2	758	1
3	of	3	481	1
4	will	4	444	1
5	to	5	442	1
6	we	6	398	1
7	in	7	277	1
8	for	8	229	1
9	a	9	158	1
10	development	10	141	1
11	be	11	132	1
12	on	12	103	1
13	our	13	83	1
14	china	14	82	1
15	people	15	81	1

图 3.9　词频统计表

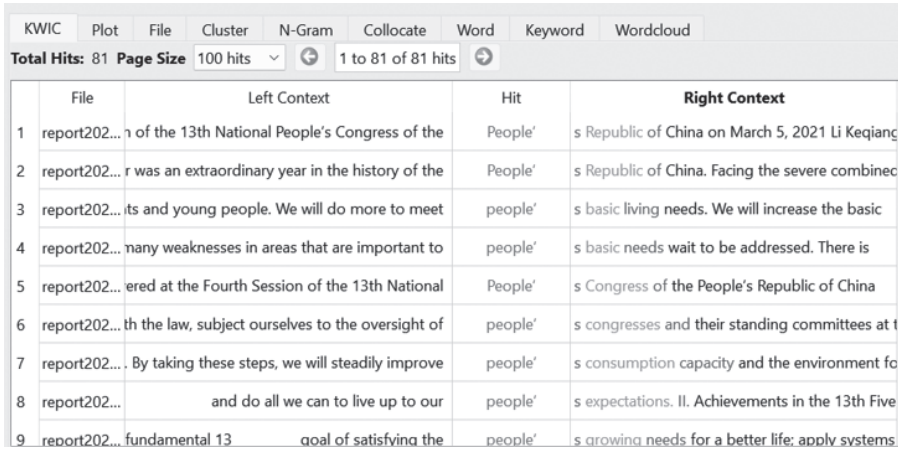
可以看到，除了虚词外，development 位居前列。这时可以使用一个英文的 Stoplist（停用词表），产出该文件的实义词词表，如图 3.10。



	Type	Rank	Freq	Range
1	development	10	141	1
2	china	14	82	1
3	people	15	81	1
4	improve	19	62	1
5	ensure	25	52	1
6	promote	25	52	1
7	rural	28	51	1
8	market	29	49	1
9	services	29	49	1
10	system	31	47	1
11	major	32	46	1

图 3.10　实义词词表（词频从高到低）

可单击相应的词，考察这个词在语境中的使用。如单击 people，在 Concordance 选项卡中就会自动出现所有包含 people 的检索结果，如图 3.11。



	File	Left Context	Hit	Right Context
1	report202...	n of the 13th National People's Congress of the	People'	s Republic of China on March 5, 2021 Li Keqiang
2	report202...	r was an extraordinary year in the history of the	People'	s Republic of China. Facing the severe combinec
3	report202...	its and young people. We will do more to meet	people'	s basic living needs. We will increase the basic
4	report202...	nany weaknesses in areas that are important to	people'	s basic needs wait to be addressed. There is
5	report202...	ered at the Fourth Session of the 13th National	People'	s Congress of the People's Republic of China
6	report202...	th the law, subject ourselves to the oversight of	people'	s congresses and their standing committees at t
7	report202...	. By taking these steps, we will steadily improve	people'	s consumption capacity and the environment fo
8	report202...	and do all we can to live up to our	people'	s expectations. II. Achievements in the 13th Five
9	report202...	fundamental 13 goal of satisfying the	people'	s growing needs for a better life; apply systems

图 3.11　people 的语境检索结果

需要注意的是，在使用 AntConc 处理中文时，使用格式与文件格式均应设置为 Unicode（UTF-8）。

除了在译文语料库中考察某一单词的用法外，双语语料库下的语境关键词检索还能为译员提供第一手材料。例如，利用 ParaConc 工具将中国银行 2007

年双语年报中的目录进行对齐（对齐过程见前文）。为了方便操作，可在导入语料之前对语料进行预处理。另外，中文需要进行分词处理才可以检索成功。例如，译员可能不熟悉“报表”在银行的年报中该如何翻译，是 Report 还是 Chart？首先，选择搜索语言为中文，输入检索项“报表”（如图 3.12）之后即可出现双语对应的语境（如图 3.13）。经过检索发现，财务报表所用的专业词汇是 Financial Statements，而非预期的 Report 或 Chart。从某种角度来说，双语对应检索的工作机制类似自建的双语词典，在实践中更能满足特定领域的个性化语境检索要求。

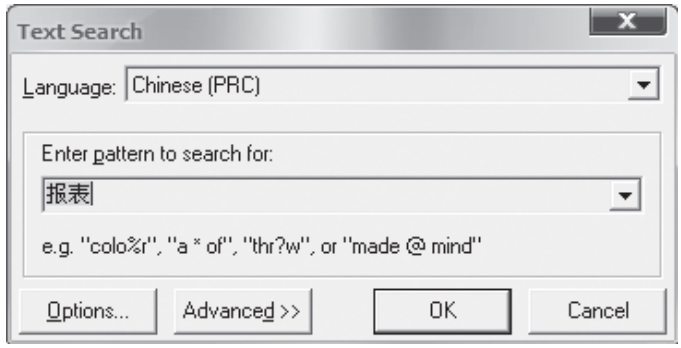


图 3.12　输入检索项

图 3.13　双语语境检索结果

另外，双语对应检索还可以帮助考察原文的特定句式（如英语中的 so ... that 或汉语中的“一……就……”）在目的语中的表达方式，这既可以作为译员培训的材料，也可以进一步进行翻译研究。王立非等（2008）早期就运用 ParaConc 进行分析，发现英语指示代词 that 在很多情况下不是被翻译为“那”，而是被翻译为“这”。

四、基于语料库的翻译研究

翻译研究是语料库应用的一个重要领域。简言之，在翻译研究领域，基于语料库的研究为认识、研究和教授翻译提供了新的方法和思路。基于语料库的翻译研究有助于在翻译描写和翻译实践之间架起一座桥梁，使翻译产品有效地服务于翻译教学、翻译理论研究和译员培训。深入的语料分析将有利于翻译等值研究、搭配或语义韵律研究等，从语料分析得到的有关句法和文本的特殊数据也对翻译实践有着积极的作用（王克非，黄立波 2007）。本节通过检索翻译英语语料库（Translational English Corpus，TEC）进行相关介绍。

TEC 在曼彻斯特大学翻译和跨文化研究中心（CTIS）初步建成，是世界上第一个翻译英语语料库。在 TEC 之前，语料收集时并没有考虑翻译文本，因为当时的研究者多崇尚“规范性”研究，认为翻译文本扭曲了规范文本，是一种偏离常态的语言。TEC 采用由欧洲和其他地区公开发行出版物的英语文本，不经过任何删改构成，源语包括法语、德语、西班牙语、葡萄牙语、意大利语、威尔士语、波兰语、阿拉伯语、汉语、泰语和希伯来语等（Olohan 2002），截至 2023 年，容量达到了 1,000 万词。

TEC 下设四个子库，涵盖杂志、新闻、小说、传记，此外还提供了一系列文本外信息，如译者的性别、国籍、职业、语言翻译方向、源语和译文的出版者等，适合进行多方面的翻译研究。

通过都柏林大学三一大学的 Saturnino Luz 博士开发的针对性工具，研究者可以从 TEC 主页上下载检索软件，免费在线浏览语料库并进行检索，[1] 如图 3.14。

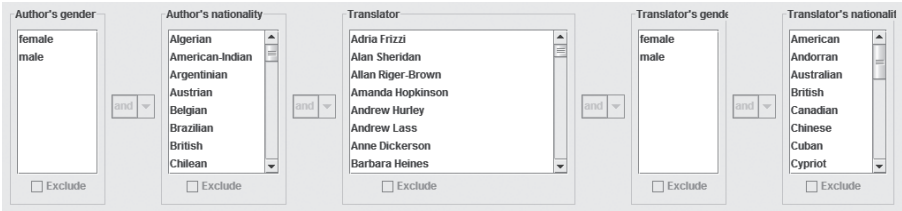


图 3.14　作者和译者信息检索

在译者下拉菜单中选择 Eliot Weinberger，再点击主菜单 Plugins 下的 Word frequency list，就可以得到该译者的译文词频表，如图 3.15 所示。

1　检索地址为 https://research.manchester.ac.uk/en/projects/translational-english-corpus-tec，2023 年 7 月 24 日。

Rank	Type	Frequency	% total
1	the	17	∞%
2	his	14	3.415%
3	he	12	2.927%
4	and	12	2.927%
5	to	11	2.683%
6	a	9	2.195%
7	don	8	1.951%
8	pedro	8	1.951%
9	of	8	1.951%
10	at	6	1.463%
11	you	6	1.463%
12	with	6	1.463%
13	in	5	1.220%
14	eyes	4	0.976%
15	upon	4	0.976%
16	hate	4	0.976%
17	was	4	0.976%
18	'	3	0.732%

图 3.15　词频表

另外，点击主菜单 Plugins 下的 Corpus description browser，还可以清晰地看到所搜集译文的详细信息，包括标记类型比（TT ratio）等（如图 3.16）。从图 3.15 的词频表可以看出，排序靠前的词除了介词和代词等虚词之外，hate 和 eyes 均出现了四次，这样的译文特征就值得进一步探究。

#	File	Subcorpus	Description	# tokens	TT ratio
1	bb000001	biography	Wittgenstein's Nephew	34,171	12.411%
2	bb000002	biography	Forbidden Territory: The Memoirs of Juan Goytisolo 1931-1956	85,636	14.579%
3	bb000003	biography	Notebooks 1924-1954, editor: Michael Tanner	79,235	8.972%
4	bb000004	biography	Realms of Strife: The Memoirs of Juan Goytisolo 1957-1982	100,137	13.607%
5	bb000005	biography	Conversations with Dvora. An Experimental Biography of the First Modern Hebrew ...	136,663	7.693%
6	bb000006	biography	Memoirs of Beethoven. From the House of the Black-Robed Spaniards, editor: Ma...	36,630	15.217%
7	bb000007	biography	The Childhood of Nivasio Dolcemare	39,540	19.170%
8	bb000008	biography	The Search. Personal Papers	39,490	12.304%
9	bb000009	biography	Story of a City. A Childhood in Amman	98,001	9.584%
10	bb000010	biography	Girls of Alexandria	63,226	15.125%
11	bb000011	biography	Delirium and Destiny. A Spaniard in Her Twenties	121,141	8.845%
12	bb000012	biography	The Boulez-Cage Correspondence, editor: Jean-Jacques Nattiez	52,042	12.417%

图 3.16　标记类型比

通过 TEC，可以进行词汇密度、词频、句子长度、搭配模式、特定词汇的使用以及使用频率的比较研究。Baker（1999）根据此语料库进行了相应的译者风格和翻译范式等研究，取得了显著成就。她总结出 TEC 能够有效揭示

译者语言使用习惯、语言行为的偏好、特殊的句法结构，以及标点符号的使用等。例如，通过研究，她发现日语文本中对翻译中外来词的忍耐程度要远远高于法语和阿拉伯语。还有学者通过 TEC 做过语义韵律的研究，考察某个单词通常同其他哪些单词一起出现，而通过总结这些搭配单词是积极意义还是消极意义，就能对原词有更深刻的了解和体会。

五、双语语料库与计算机辅助翻译

除了检索并考察语境关键词之外，双语语料库同计算机辅助翻译软件结合起来，就可以形成翻译记忆，协助完成翻译任务。Bowker（2002）曾简单明了地将翻译记忆解释为“一组语言文本的句子与其在目标语中相对应的句子”。翻译记忆的作用原理是用已有译文建立双语语料库作为记忆库，通过浏览已经翻译好的文本，在进行同原文相近的新任务时，提取翻译记忆进行提示和替换，通过寻找形式上的相似性协助完成翻译任务。对此，本书在后续章节有详细说明。

翻译记忆是针对句子或篇章层面的对等，而由专业术语形成的双语术语库则保证了翻译任务的术语统一。主流的计算机辅助翻译软件都配有团队工作工具，允许不同的译员共同合作完成一个翻译项目。如果没有双语术语库为翻译记忆系统统一术语，那么译员合作的结果就不能保证高度精准的术语统一。除了术语统一，还要实现精确的语言转换。双语语料库形成的翻译记忆有助于保证译文与历史译文资料统一、与指定文献统一等。在翻译项目进行的过程中，译员形成的新语料又被不断地纳入语料库中，团队内的其他人就可以通过更新语料库来使用同样的句式和表达，而不用再重复翻译。

利用双语语料库，在翻译项目中能同步提取文章中出现的术语，经专家质量检验后，按照不同领域进行区分，更新为新的双语术语库。转换为翻译记忆之后，这些术语就能在以后的翻译项目进行时再利用。相对于此，传统意义上人工积攒的术语库，不易分类查找，不能及时更新，且数量有限，已经不能满足高效率翻译实践的需求。

柯飞（2002）曾将自动翻译过程简要地总结为四步：第一步，将双语语料平行对齐；第二步，对语料给予相关联的标注；第三步，将汉语作分词处理，并根据词频计算权重；第四步，通过权值和字串对比，计算和检索跟使用者输入的文字相对应的句子并显示出来。

Hunston（2002）指出，基于平行语料库的机器翻译系统通过对短语而不

是对词的识别可以使机器翻译更加准确，包含不同语言的平行语料库对于译员来说用处更大。因此，对那些翻译活动占据重要地位的组织和机构来说，进行平行语料库同机器翻译结合的研究就十分重要。欧盟就在不断改进自动翻译过程。例如，由 John Sinclair 与 Wolfgang Teubert 负责协调的跨欧洲语言资源建设学会（Trans-European Language Resources Infrastructure，TELRI）曾经出版了 CD-ROM 格式的语料库材料（Erjavec *et al*. 1998），包含有柏拉图《理想国》多语语料库（包括 17 种欧洲语言的译文）等，每种语言都同原文进行了句子层次的对应。1977 年，加拿大蒙特利尔大学研发的自动翻译系统 TAUM-METEO 能在一天内把加拿大各地区的气象预报从英语翻译成法语，这标志着第一代机器翻译系统的诞生。在我国，研究自动翻译的机构和项目也有很多，如北京大学计算语言学研究所、清华大学智能技术与系统国家重点实验室和中国科学院计算技术研究所联合承担了国家 973 课题“面向新闻领域的汉英机器翻译系统”，以及中国科学院自动化研究所模式识别国家重点实验室研制的口语自动翻译系统等。

除了使用对齐的双语语料库改进机器自动翻译的质量，黄俊红等（2004）认为，双语语料库还可以加强机器辅助翻译中的人机交互，通过统计模型从双语语料库中获取翻译模型，从而改进费时、易出错的传统机器翻译模型。也有学者认为，基于语料库的机器翻译系统能够大大超过第三代机器翻译系统的性能，很可能成为第四代机器翻译系统的雏形（李亮 2004）。

自 20 世纪末人文社科研究领域进入数字人文（digital humanities）时代，计算机技术与语言研究和翻译研究的结合更为紧密和复杂。数字人文是结合了人文学科和数字技术的交叉学科。它主要探究人文学科领域内数据、信息和知识的数字化、组织和使用。它通过技术手段促进人文学科研究，包括但不限于人类语言、文学、艺术、历史、文化学、哲学等方面的研究。在英语语言研究领域出现了不少针对性工具和研究成果，如美国卡内基梅隆大学 David Kaufer 主持开发的 DocuScope 英语文本修辞功能分析工具及其一系列成果。在涉及中文的翻译研究领域，Qian & Kaufer (2017) 利用 DocuScope 以及计算机辅助的人工分析，对中国的政府工作报告进行了双语的修辞功能分析。胡开宝和王晓莉（2022）在指出数字人文研究现状的不足之后，提出应建设并运用各种翻译研究数据库，并将原生性数字文本纳入该领域的研究对象之中，同时也应当根据研究目的选用不同的数字人文研究方法，依据数字人文和翻译学的相关理论构建并完善该领域研究的理论框架。耿强和周知非（2023）利用计算机辅助工具，考察《人民日报》1949—1966 年间所生产的中国翻译话语特点，发现了此话语由信仰和技术所构成的深层二维结构，帮助深化了对中国当代翻译话

语的再认识。李崇华和张政（2023）提出了数字人文视域下的改进方案，如基于文献计量学工具和语料库方法完善理论体系，构建大型、共享、高质量的数据库用以提升规范性，通过 Python 和 R 等计算机编程语言研究接受状况等。

六、小结

本章从双语语料库的概念和类型划分谈起，着重介绍双语语料库的建设步骤、国内外双语语料库的建设情况等，并简单提及双语语料库在翻译实践中的作用。

目前国际上建立了各种类型的语料库，不同类型语料库可为译员和校对人员提供大量语言实例。双语语料库可分为三类：平行语料库（包括单向对应、双向对应、多向对应语料库）、多语语料库、可比语料库。双语语料库建设涉及确定建库目的、搜集双语材料、语料标注（同实践相关的主要是中文切分）、双语语料库对齐、检索软件开发与后期库维护等内容，与计算机辅助翻译工具的研发密切相关。[1]

本章还对几个较有名的平行语料库予以举例，说明双语语料库在检索方面的应用，并简要介绍了双语语料库的广泛应用，如翻译研究、语境检索、翻译教学、计算机辅助翻译、机器翻译等。可以看出，双语语料库与翻译技术是密切联系，相互促进的。

展望未来，数字人文技术在翻译研究、翻译教学、翻译实践中的应用将会越来越广泛。首先，数字人文可以为翻译研究提供数据的处理和分析支持。例如，使用自然语言处理技术，计算机工具可以帮助翻译研究者进行语言文本的分析和挖掘，发现不同语言文本之间的异同，辅助了解跨文化差异。其次，数字人文可以为翻译研究提供信息和文献的收集和管理工具，帮助翻译研究者快速获取相关文献和信息，识别出有用的信息和数据，提高工作效率。数字人文还可以帮助翻译研究者进行翻译记忆库的构建和管理，帮助建立更大规模、更全面的翻译记忆库，为翻译工作提供更多实用的数据和信息。最后，它可以为翻译研究提供各种数据可视化和探索工具，通过图形化的展示和呈现提高翻译研究的可视化效果，更全面、更直观地呈现出翻译工作的结果。

正如张威和雷璇（2023）所指出，面对数字人文这一时代热潮，要全面评价数字人文语境中的翻译研究现状及特征，须客观分析在本体观照、理性分

1　需要注意的是，语料库的建设与使用都耗时耗力，受到外界客观因素（如资金和人员变动）的影响很大，一些网站或网页的功能会因此受到限制。

析、实际应用等方面的问题，以明确数字人文性质翻译研究独特的跨学科属性及人文学科价值与定位。

拓展阅读推荐

Kennedy, G. (2014). *An Introduction to Corpus Linguistics*. London: Routledge.

此书介绍了语料库语言学的基本概念、方法和应用，为如何构建和利用语料库提供了详细指导，并探讨了语料库分析在各个领域的应用。

Qian, D. & Kaufer, D. (2017). A rhetorical approach to translation: The Chinese "Report on the Work of the Government" as a case study. *Translation Spaces*, 6(2): 270-290.

耿强，周知非．数字人文视域下《人民日报》（1949—1966）生产的中国翻译话语研究．外语电化教学，2023 年第 1 期．

胡开宝，王晓莉．数字人文视域下翻译研究：现状、问题与前景．外语与外语教学，2022 年第 6 期．

李崇华，张政．基于数字人文的视听翻译研究范式与理路．外语学刊，2023 年第 1 期．

秦洪武．双语语料库的研制与应用．北京：外语教学与研究出版社，2021．

此书介绍语料库建库理念、制作过程和应用工具，并梳理经验性语言数据的分析和统计方法，可供译者、翻译学习者、翻译研究者和语言研究者有效运用语料库从事翻译、翻译教学和对比语言研究使用。

王克非．双语对应语料库：研制与应用．北京：外语教学与研究出版社，2004．

此书介绍了双语对应语料库的研制及其相关研究，如双语对应语料库检索研究、基于对应语料库的语言研究和基于对应语料库的翻译研究。

张威，雷璇．翻译研究的数字人文“转向”：现状及反思．中国翻译，2023 年第 2 期．

思考与练习

1. 把自己平时翻译实践所产出的双语文本进行对齐，检索任意高频关键词，观察是否对应，并思考这说明了翻译过程中的什么情况？
2. 下载 2022 年中国《政府工作报告》的中英文版本，按照本章所介绍的方法处理文本，导入 ParaConc 创建小型双语语料库，导出汉英双语词频表数据，并观察数据是否与 AntConc 导出的数据相同。
3. 使用本章提供的在线双语语料库资源，查找自己感兴趣的关键词，观察其使用情况是否与自己的预测情况相符。