第三章

如何选择合适的神经机器翻译解决方案：机器翻译质量评测

Caroline Rossi
格勒诺布尔-阿尔卑斯大学

Alice Carré
格勒诺布尔-阿尔卑斯大学

机器翻译发展迅速，但无法提供放之四海而皆准的解决方案。为了给特定项目选择恰当的解决方案，用户需要进行多种比较和评测。这从来都并非易事，尤其当机器翻译输出的质量看似越来越好，错误越来越难识别。我们如何才能最恰当地定义和评测神经机器翻译解决方案的质量，从而做出正确的选择呢？第一步当然是尽可能明确需求。本章先从语用角度给质量下定义，然后介绍人工评测和自动评测机器翻译质量的关键概念，并概述译者可以如何运用这些概念。

1 引言

神经机器翻译（NMT）被大肆宣传的同时，用户也的确注意到了机器翻译输出质量的提升。本章旨在说明，机器翻译输出虽看似更流畅，但并不一定更易处理。此外，神经机器翻译的输出译文总在变化之中，其使用应考虑具体情况和最终用户的需求。接下来，我们会提出了对质量和衡量标准的定义，这些定义和衡量标准并不局限于神经机器翻译输出表面上的易用性和流利度。

本章首先要回答的问题是：如何以可靠且有效的方式来评测神经机器翻译解决方案？根据使用场景和文本类型，答案可能会有所不同。接下来，我们将介绍机器翻译评测的关键问题，希望能帮助用户选择适合其特定需求的机器翻译引擎。

Caroline Rossi & Alice Carré. 2025. 如何选择合适的神经机器翻译解决方案：机器翻译质量评测. In Dorothy Kenny (ed.), 机器翻译知识普及：为人工智能时代的用户赋能, 39–63. Berlin: Language Science Press. DOI: 10.5281/zenodo.14922289

1.1 何为机器翻译质量评测?

广义的翻译质量包括产品和过程:“翻译质量既包括最终产品(译文)的质量,也包括交易(所提供服务)的质量”(Gouadec 2010: 270)。此外,翻译质量评测在很大程度上取决于完成翻译时的场景,而且翻译教师和有特定使用目的的客户对翻译质量的期望肯定有所不同。换句话说,“质量的概念是相对的”(Grbić 2008: 232)。在翻译研究领域,翻译质量众所周知地难以定义且千变万化。因此,许多综述研究便根据不同的翻译理论来定义翻译质量(参见例如Drugan 2013、House 2015)。在机器翻译质量评测中,也存在对“质量”的不同定义。“翻译技术(特别是机器翻译)的发展和广泛运用,催生出太多不明晰且操作方法不同的质量的定义及其相应的衡量标准”(Doherty 2017: 131)。就机器翻译而言,质量评测往往被视作实现某个目标(即改进系统)的手段,因此语用维度的考量便占据主导地位,通常结合人工评测和自动评测(Doherty 2017: 133)。在介绍现有的评测方法以及如何将其结合使用之前,我们先来解释语用维度在机器翻译质量评测中的必要性。

对于机器翻译的用户来说,机器翻译输出的质量评测过于复杂。虽然输出质量主要由系统本身决定,但翻译场景和用户需求也是质量评测需要考量的关键因素。举个简单的例子:你或许觉得适应用户手册中翻译不恰当的使用说明并不难,因为你已经清楚知道如何使用自己购买的产品,或者图示说明就足够了。在这种情况下,我们很可能对机器翻译错误有相对较高的容忍度(Castilho & O'Brien 2016)。现在,我们切换到另一个完全不同的场景,也涉及技术文本,但增加了法律层面的考量:专利证书译文的使用者需要获得精确的相关信息,因此对机器翻译错误的容忍度会低得多。提及专利领域的神经机器翻译,Castilho et al. (2017: 113) 的研究表明,神经机器翻译时常漏译原文信息,而在这种翻译场景中的机器译文信息缺失会造成严重后果。在上述两个例子中,质量评测的语用维度考量也就意味着采取可测量的有用性指标,如用户满意度评分、译后编辑效率的提高或机器翻译的产品描述带来的销售额增幅。

总的来说,翻译质量评测绝非小事。无论是人工评测还是自动评测,都会受到多重因素的影响。首先,合理的译文通常并非只有一个,也就是说原文存在多种同样可以接受的译法。此外,人工评测往往比较主观——事实上,评测者对统一译文质量水平存在不同看法的情况并不少见。但是,基于评测者处理译文的过程来计算评测效率,却不失为一种的客观方法。总之,为了减少主观影响,我们必须明确界定评测的目标和指标。人工评测的另一个缺点是过程耗时且占用大量资源。作为人工评测的替代方案,基于算法的自动评测尽管肯定比人工评测的成本更低、速度更快,但有时关联度低,因为自动评测无法评估译文在具体应用场景的可用性。因此,这两类评测方法各有利弊,应根据翻译项目和需求来进行选择。

1.2 质量恰好：得再三考虑！

如上所述，若从语用维度来进行机器翻译质量评测，那么显然并非所有错误的严重程度都一样。我们常以食谱翻译为例，因为译文质量很容易检验（“尝”得出来）。其实，食谱还常被用来展示机器翻译如何闹笑话。例如，以下为用机器将法语译成英语的千层面食谱，试问，单看译文，你能做出千层面吗？

表 1: 食谱译文的神经机器翻译错误示例（划线部分）(法译英)

法语原文[a]	英语译文（神经机器翻译）
Préchauffez votre four à 180°C. Dans un plat à gratin beurré, versez un peu de béchamel. Déposez une couche de pâte et poser une couche de farce a la viande. Déposez à nouveau des pâtes, béchamel et viande.	Preheat your oven to 180°C. In a buttered gratin dish, pour a little béchamel. Put a layer of dough and a layer of meat stuffing. Put a layer of pasta, béchamel and meat filling on top.
Terminer par une couche de pâte avec de la béchamel et saupoudrez de fromage râpé. Laissez cuire 35 à 40 minutes.	Finish with a layer of dough with béchamel and sprinkle with grated cheese. Leave to cook for 35 to 40 minutes.

[a]奶奶千层面（法语食谱）: https://www.750g.com/lasagnes-r66998.htm

你想必会惊讶地发现，原文中 pâte（面团；面条）出现了单数和复数形式，但均指意大利面，而机器翻译将单数形式译为面团。聪明谨慎如你或许能猜到正确词义，但一味按照通顺的机器翻译食谱来操作，你可能会做成另一道菜，而问题仅仅源于一个字母（表示复数的词尾）。尽管机器翻译译文总体流畅准确，稍作修改即可提升质量，但只要存在一个严重问题，就会让译文的功能失效。食谱内容相对简单，再加上如果对常见菜肴有所了解的话，读者可以自行纠正译文错误，理解起来并不困难。但是，对于更复杂的其他文本类型或专业领域，缺乏背景知识的读者就没这么好理解了。另外，这里提出的评测方法涉及到人、配料和厨房：测试成本或许很高，且几乎从未被使用过。

除了这类错误，大多数机器翻译用户很可能会遇到抽象概念和隐喻表达的翻译问题。表 2展示了一本法语原著的简介摘录，以及神经机器翻译系统输出的英语译文。请问英语读者知不知道 a veritable pie in the sky（法语原文为 véritable tarte á la crème）意为 well-trodden path（常有人走的路），还是 prefabricated subject（预制对象）？

表 2: 习惯用语的误译示例（带下划线）

法语原文[a]	英语译文（神经机器翻译）
L’ indépendance du parquet, véritable tarte à la crème remise sur le plateau à chaque campagne présidentielle, était aussi une proposition du candidat Macron. Il ne l’ a pas tenue.	The independence of the public prosecutor’s office, a veritable pie in the sky put forward in every presidential campaign, was also a proposal by candidate Macron. He did not keep it.

[a]法语原文来源: https://www.grasset.fr/livres/ministere-de-linjustice-9782246827504

你如果经常使用机器翻译，或许能识别出上述或其他的翻译错误。经验很重要！机器翻译译文越是通顺流畅，就越需要谨慎对待——最近有研究表明，学生对神经机器翻译译文的修改率低于流畅度较差的其他类型的机器翻译系统 (Yamada 2019)。熟悉某个语言对或领域的神经机器翻译译文的常见错误，有助于使用者更有效地识别和改正这些错误。

总之，尽管神经机器翻译的质量确实越来越高，这点无可否认，但它可能并不比其他类型的机器翻译更容易处理，而机器翻译也从来都不是唾手可得的灵丹妙药。相反，我们需要注意到隐藏在流畅的机器翻译译文中的微小错误，并在选择合适的机器翻译解决方案前，仔细思考自身需求。

2 为专业翻译选择合适的机器翻译引擎

上一节主要概述了神经机器翻译的一般使用和评测，而为专业翻译选择合适的机器翻译引擎则意味着要考虑很多其他因素。关键在于，适合专业用途的机器翻译引擎不仅必须满足客户的隐私要求、完美融入工作流程、支持所需语言对，还能输出只需最小的译后编辑工作量，就能提供满足客户需求的译文。机器翻译输出的质量将取决于专业领域和文本类型、引擎的可训练性以及译前编辑和译后编辑的工作量。

2.1 隐私和保密

机器翻译系统是如何处理用户数据的？尽管这个问题对你、你的业务员或客户来说都至关重要，但大多数机器翻译解决方案并没有给出明确回答，或没有在最开始就隐私问题发出警示。相反，你必须仔细阅读隐私声明，并确保在使用前做出正确选择。

例如，欧盟的神经机器翻译系统 eTranslation 的隐私声明中警示道：“用户向任何在线服务（包括 eTranslation）提交潜在敏感文件时，应自行做出判

断”。即使数据不会被保存在机器翻译供应商的服务器中，这类警示也是正当且值得关注的。其实，就算已上传的数据不久后便会删除，也有可能造成隐私问题。出于保密考虑，欧盟法院的听证会甚至禁止使用 eTranslation 等内部机器翻译系统，因为没有哪个机器翻译系统能够达到此类听证会的严格保密要求[1] (C. Lenglet，个人交流)。

当然，免费在线机器翻译存在的隐私问题风险要高得多，因为数据会被一直保存和重复使用。因此，无意中输入免费机器翻译系统的文本的某些机密信息，可能会以意想不到的方式泄露（请参阅Moorkens 2025 [本卷]深入了解如何合乎道德地、安全地处理数据）。

2.2 比较机器翻译输出

你或许想用同一篇原文来测试几种机器翻译工具，然后通过比较其输出以判断哪种最符合自己的需求。

根据你想解决的问题，机器翻译输出的比较方法有以下几种。例如，可以通过人工评估、自动评测和/或译后编辑工作量给译文打分，然后比较得分情况。译后编辑工作量的计算方法详见O'Brien (2025 [本卷])。可以说，译后编辑工作量小或不需要进行译后编辑的机器翻译输出，一般认为比译后编辑工作量大的“更好”；还可以说，如果合理推断机器翻译文本会用于传播目的，则译后编辑工作量可用于评测译文质量（可参阅Kenny (2025 [本卷])详细了解以同化和以传播为目的的机器翻译之间的区别）。我们将在下文重点介绍机器翻译输出的人工评测和的自动评测。

2.3 人工评测

人工评测需要依靠评测者来评估一个或多个系统的输出，通常逐句（或逐个语段）进行，但也可以进行文档级别的评估（例子可参见Castilho 2020）。评测者通常从两个方面来评分。第一是忠实度 *(adequacy)*，即衡量机器翻译译文对原文意思表达的完整程度。忠实度的衡量通常采用范围从 1（指机器翻译完全没有表达出原文含义）到 4（机器翻译完整表达出原文含义）的四级量表。有时也使用五级量表，不过由于量表级别数为单数，评测者可能会过度倾向于选择中间项。第二是流利度 *(fluency)*，即衡量“译文在多大程度上遵循了目的语的规则和规范” (Castilho et al. 2018: 18)。理论上，评测者对流利度的判断无需借助原文。其测量通常也使用四级量表，1 表示机器翻译译文“不流畅”，4 则表示该译文的流利程度达到母语水平 (Castilho 2020: 1152-1153)。忠实度和流利度的评测非常耗时，因此操作成本很高。

另一个简单快捷的机器翻译系统输出的比较方法是，只需评测者对译文的质量进行排序，也就是说，判断哪种“更好”即可，无需说明理由。目前已

[1]详细描述和解释可参考：https://eur-lex.europa.eu/legal-content/en/TXT/PDF/?uri=CELEX:32013D0488

有很多机器翻译供应商采用这种方法，以便迅速获得在线用户的反馈。例如，Moorkens (2018)指出，2017 年，微软采用该方法获得了用户对其统计机器翻译和神经机器翻译系统输出的评测结果。

其他机器翻译供应商已经开发出更复杂的工具来协助机器翻译输出的人工评测。例如，Kantan AI 提供了名为 KantanLQR（Language Quality Review，即语言质量审查）的工具，允许用户按需自行选择质量评测标准（如忠实度、流利度和术语使用等），然后根据这些标准来比较多达四种不同的机器翻译输出。[2] 这类工具非常实用，不仅能通过饼图和条形图为评测者对每个句段的质量得分进行视觉化呈现，还能计算不同机器翻译引擎或系统的总体得分。此外，这类工具还具备为翻译公司的项目经理以及评测者而设计的功能。PET(Aziz et al. 2012) 等非商业工具也可用于帮助评测者评估机器翻译输出，同时备受研究人员的青睐。

其他支持人工评测的常用工具还包括电子表格程序，使用者可以先将分数手动输入表格，如表 3所示，然后用内置函数来计算质量指标的平均得分。各种免费在线表格也可以用来进行人工评测。[3] 线上表单尤其适用于调查问卷，也和电子表格一样能够自动计算结果以及其他统计学指标。

表 3: 用于比较机器翻译系统的电子表格示例

原文	NMT 译文 1	NMT 译文 2	NMT 译文 3	偏好译文	注解
语段 1					
语段 2					

2.4 错误类型

有时，评估者不仅要使用上述标准对机器翻译的段落或文档进行评分，还要根据错误分类法 *(error typology)* 将译文中的每个错误归类，具体说明译文的问题所在。对错误进行分类是识别机器翻译输出问题的重要步骤，通常是为了向系统开发人员提供反馈意见。

但是，错误类型往往非常复杂。例如，多维度质量评估（Multidimensional Quality Metrics, MQM）框架列出了详尽的错误类型 (Mariana et al. 2015: 140)。为便于操作，评测时可选择其中的常见类型，例如 Moorkens 为课堂翻译评测练习选择的错误类型 (Moorkens 2018: 380) 包括：

- 词序错误（短语或单词级别的词序不正确）

[2]详见 https://kantanmt.zendesk.com/hc/en-us/articles/115003644483-What-is-KantanLQR- 和 https://twitter.com/i/status/1466392446552657927.

[3]最有名的或许要数谷歌表单（Google Sheet）。可参见 https://support.google.com/docs/answer/6281888?hl=en&co=GENIE.Platform%3DDesktop

- 误译（单词、性别、数字或大小写的翻译有误）
- 漏译（原文中的单词在目标文本中没有译出）
- 增译（原文中没有的词却在目标文本中译出）

有可能用于评测机器翻译输出的样本太小，不足以说明每个机器翻译引擎的水平。理想的情况应该是，用不同的样本进行重复比较，然后找出性能最佳的机器翻译引擎或系统。不过，大型机构可能有条件进行大规模评测，而较小的翻译服务提供者和自由译者可能更倾向于采取自动评测和计算译后编辑工作量的方法。

2.5 自动评测指标的使用

自动评测指标（automatic evaluation metrics，AEM）比人工评测的速度更快且成本更低，因此机器翻译用户可根据需要频繁对机器翻译输出进行评测。例如，在机器翻译引擎的训练过程中，用户可以每调整一次参数便进行评测，看看引擎的性能是否有所改善。如果不是训练引擎，而是为某个项目选择一个合适的引擎，那么自动评测指标也能用于评测同一原文的多种机器翻译输出。

对于同一原文，人工翻译的译文经常是千差万别。因此，我们不能指望机器翻译系统与人工翻译达到完全一致。但是，与人工翻译越相似，机器翻译的质量可能越好，反之则越差，因此许多自动评测指标都是基于相似性原则，即向评测工具同时输入人工翻译作为“黄金标准”（或称“参考译文”）和系统输出（或称“候选译文”，有时也称为“假设 (hypothesis)”）。然后，将候选译文与参考译文进行对比，并计算相似性（similarity）或相异性（dissimilarity）。考虑到会存在多种参考译文，一些评测工具甚至可以采用多种参考译文。[4]

过去的二十年里出现了大量自动评测指标及其变体。不过在此，我们只会根据章节内容会提到的评测工具，如 KantanMT 和 MutNMT，有选择性地为读者介绍几种自动评测指标。想了解更多有关自动评测指标信息的读者，可参阅Koehn (2010)和Koehn (2010)。为保持一致，我们将采用 Koehn 提出的术语和标记方法，并在适用的情况下，在以下例子中运用 Koehn 的解释来评测神经机器翻译的输出。

以下例子摘录自一种无线电收发器的用户手册[5]，如图 1所示。

手册中逐点列举了确保该收发器防水的前提条件，上述原文便是其中一点 (参见 图 2)。

[4]在这种情况下，需要决定如何计算参考译文的长度。例如，同时采用多种参考译文的 BLEU（双语替换评测）采用的是“长度最接近候选译文”的参考译文。(Qin & Specia 2015: 114)

[5]Vertex Standard 的 VX-450 系列现已停产。

原文：	Battery pack is attached to the transceiver.

图 1: 本节主要示例的原文

重要提示
只有在下列条件下，才能保证收发器的防水性能（IP57：1 米/30 分钟）(*sic*)：

- 电池组附着于收发器；
- 天线连接到天线插孔；
- MIC/SP 帽安装在 MIC/SP 插孔中。

图 2: 收发器用户手册摘录

虽然该手册是面向大众而撰写的，但原文属于技术文本，因此其翻译可视为专业领域翻译。文本涉及无线电通信领域，因此必须采用该领域的术语和措辞，而作为用户手册，译文也应保留这类体裁的特点。例如，每个概念都应只用一个术语来表示（即不应使用同义词），且每个术语都只对应一个概念，即“单义性”（monosemy）。手册说明应简洁易懂，写作模式统一。（更多关于专业领域和文本体裁的介绍，请参见Kenny 2025 [本卷]。）在我们上述的翻译任务中，目标语是法语，译文用途与原文的一致，即供收发器的用户使用。

接下来，我们就这段摘录为原文，对比三种机器翻译工具的输出。首先是欧盟的机器翻译系统 eTranslation（以下简称“系统 A”，其输出简称“候选译文 A”）。[6] 其次是谷歌翻译（以下简称“系统 B”，其输出简称“候选译文 B”）。[7] 最后是 DeepL（以下简称“系统 D”，其输出简称“候选译文 D”）。[8] 在撰写本书时，这些系统均对公众免费开放，不过 eTranslation 要求用户进行注册，且身份类别为中小企业、公共服务官员或公共部门服务提供商。

本章节介绍的自动评测指标中，有些简单方法可以手动计算。但对于复杂的方法，我们会借助 MutNMT 来计算。[9] 我们会向读者举例，帮助了解这些指标的运算过程，但还想说明的是，分数的精确计算会根据评测指标的具体操作细节有所不同———如果使用不同的工具来计算似乎相同的 AEM（例如 BLEU），您很可能会得到不同的结果。[10] 评测结果出现差异的原因可能有

[6] https://webgate.ec.europa.eu/etranslation/translateTextSnippet.html

[7] https://translate.google.com/?hl=en

[8] https://www.deepl.com/en/translator. 请注意，我们将 DeepL 称为“系统 D”，而不是“系统 C”，是为了避免在使用字母 C 来指代候选译文 (candidate translation) 的情况下出现混淆。

[9] https://mutnmt.prompsit.com/index

[10] 不过，最近有开发者对努力实现自动评测指标的规范化和参考应用，如 Matt Post 的 sacrebleu 等软件 (Post, 2018)。

以下几个：评测工具在计算前处理引号、连字符、普通空格和不断行空格等的方式不同；对于词符（token）的定义方式不同（是否考虑到撇号、连字符、标点符号或语言学信息，如词目 (lemmas) 或多词单位）；对大小写是否敏感；度量参数细节（例如，具体操作中 *n*-grams 的顺序）。[11] 在上述例子中，我们将候选译文 D 中的撇号改为参考译文所使用的撇号。这样一来，自动评测指标的结果就不会受到不同撇号的影响，我们便可以只关注机器翻译输出本身。此外，我们为了展示而手动计算自动评测指标时，将连字符和撇号视为单词“分隔符”，也就是说参考译文（参见图 3）共包含 8 个词。

参考译文：	La batterie est installée sur l’émetteur-récepteur.

图 3: 参考译文示例

图 4 列出了我们将在下文讨论的原文、候选译文和参考译文。

原文：	Battery pack is attached to the transceiver.
参考译文：	La batterie est installée sur l’émetteur-récepteur.
候选译文 A(eTranslation)：	Le bloc-batterie est fixé à l’émetteur-récepteur.
候选译文 B(谷歌翻译)：	La batterie est fixée à l’émetteur-récepteur.
候选译文 D(DeepL)：	Le bloc-piles est fixé à l’émetteur-récepteur.

图 4: 本节的主要示例——原文、参考译文和候选译文

对于上述例子，人工评测者会给出什么反馈呢？第一点，除非客户另有说明，术语 battery pack（电池组）应译为 batterie。而候选译文 D 中译成 bloc-piles 显然是错的，因为这种收发器使用的并非 piles（不可充电的一次性电化学电池），而是 batterie（可充电的电池组）。在这种情况下，收发器使用的是锂电池。候选译文 A 中译作 bloc-batterie 本质上没错，但这不符合译入语表达习惯，而是仿造词 (calque)，即对英语原文的逐字翻译。第二点，动词词组 is attached to 可以根据个人喜好译为 est installée sur 或 est fixée à。第三点，transceiver（收发器）是 transmitter-receiver（发射接收机）的缩写，理想情况下应译为 émetteur-récepteur，如图 4中列出的所有译文所示。不过，对这个翻译任务来说，译作 radio（无线电），甚至 appareil（设备）也是可以的（关于“翻译”和“对等”的更多内容，请参见Kenny 2025 [本卷]）。现在，让我们深入了解自动评测指标是如何评测候选译文的。

[11]本书作者要对 Gema Ramírez-Sánchez 的解释致谢。

2.5.1 核心概念：*n*-grams、精确率、召回率和 F 值

我们将在本节介绍四个概念：*n-grams*, 精确率, 召回率 and *F* 值。这些概念有助于我们理解下文涉及的复杂的自动评测指标。

2.5.1.1 *n*-grams

在翻译领域，*n*-grams（参见Kenny 2025 [本卷]）通常被理解为 n 词序列。如上述例子中，battery 是 1-gram 或一元分词 (unigram)，battery pack 是 2-gram 或二元分词 (bigram)，battery pack is 是 3-gram 或三元分词 (trigram)。此外，还有 4-gram、5-gram 等，如 battery pack is attached 为 4-gram。

n-grams 通常用于语言建模，例如，一个 3-gram 的概率表示假设已知某个词前面的两个词，则该词出现的概率。

在自动评测指标中，*n-grams* 指的是参考译文中与候选译文匹配的 n 词序列。近来，有研究者提出了基于字符序列，而非词序列的自动评测指标。那么，N-grams 就可以理解为 *n* 个字符的序列，而非 *n* 个单词的序列。

讨论 BLEU 时，*n-grams* 指的是 *n* 词序列（参见 see 2.5.4.），而讨论 ChrF3 时，*n-grams* 指的是 *n* 字符序列 (参见2.5.5)。

2.5.1.2 精确率与召回率

精确率是自然语言处理许多分支都会用到的基本概念。我们借用一个简单的例子来解释：假设老师让学生用英语说出一个星期的每一天，学生回答“星期一和星期二”，则他给出了两个正确答案，没有错误答案。由于“精确率”指的是给出的正确答案与答案总数之比，因此该学生的得分是 2/2，即精确率达到惊人的 100%。

但了解正确答案的老师知道该学生的回答并不完整，因为他漏掉了其他五天。因此，老师可以说该学生回答的“召回率”低。“召回率”指的是给出的正确答案与正确答案总数（理想答案）之比。在这种情况下，该学生回答的召回率为 2/7，相当于略低于 29%。

对机器翻译译文进行自动评测时，精确率为候选译文中的正确词数（即同时出现在参考译文中的词）与其总词数之比：

$$C\text{的精确率} = \frac{C\text{的正确词数}}{C\text{的总词数}} \tag{1}$$

C 表示候选译文。

一起来看看上述的例子。在图 5中, 不同的候选译文对比参考译文。“正确”用词 (即同时出现在参考译文的词) 用下划线标出, 而“错误”用词 (即没有出现在参考译文的词) 则用删除线标出。

原文:	Battery pack is attached to the transceiver.
候选译文 A(eTranslation):	~~Le bloc~~batterie est ~~fixé à~~ l’émetteur-récepteur.
候选译文 B(Google Translate):	La batterie est ~~fixée à~~ l’émetteur-récepteur.
候选译文 D(DeepL):	~~Le bloc-piles~~ est ~~fixé à~~ l’émetteur-récepteur.
参考译文:	La batterie est installée sur l’émetteur-récepteur.

图 5: 原文、参考译文和候选译文

现在来计算每个候选译文的精确率。候选译文 A 的正确用词为 5/9, 即精确率为 0.56 或 56%。[12]候选译文 B 的正确用词为 6/8, 即精确率为 0.75 或 75%。最后, 候选译文 D 的正确用词为 4/9, 即精确率为 0.44 或 44%。由以上结果得出, 候选译文 B 优于候选译文 A 和候选译文 D。

同理, “召回率”指的是候选译文的正确词数与参考译文的总词数之比:

$$C\text{的召回率} = \frac{C\text{的正确词数}}{R\text{的总词数}} \tag{2}$$

C 表示候选译文, R 表示参考译文。

换句话说, 召回率不仅考虑候选译文用了哪些词, 还考虑应该用哪些词。

回到以上的例子 (图 6)。

原文:	Battery pack is attached to the transceiver.
候选译文 A(eTranslation):	~~Le bloc~~-batterie est ~~fixé à~~ l’émetteur-récepteur.
候选译文 B(Google Translate):	La batterie est ~~fixée à~~ l’émetteur-récepteur.
候选译文 D(DeepL):	~~Le bloc-piles~~ est ~~fixé à~~ l’émetteur-récepteur.
参考译文:	La batterie est installée sur l’émetteur-récepteur.

图 6: 原文、参考译文和候选译文

[12]在本节中, 所有在 0 到 1 区间的结果将保留两位小数, 而百分比则保留两位数。

参考译文共有 8 个词。候选译文 A 有 5 个词出现在参考译文中，则召回率为 0.63 或 63%。候选译文 B 有 6 个词出现在参考译文中，则召回率为 0.75 或 75%，而候选译文 D 有 4 个词出现在参考译文中，则召回率为 0.5 或 50%。根据召回率的计算结果，候选译文 B 依然优于候选译文 A 和候选译文 D。

2.5.1.3 *F* 值

上述例子中的学生可优先考虑精确率，只回答“星期一、星期二”即可，而不必冒险给出错误答案；也可能优先考虑召回率，一口气说出几十个答案，希望能尽量命中正确答案。所以学生可能会回答“星期一、星期二、星期三、星期四、星期五、星期六、星期日、一月、二月、三月、四月、五月、六月、七月、八月、九月、十月、十一月、十二月”。此时，召回率突然升到 100%，因为他们的回答包括了所有正确答案；但准确率竟不足 37%，因为在 19 个回答中，只有 7 个是正确的。在老师看来，这两种策略都不可取。老师想要的是，让学生同时最大化精确率和召回率。他们需要一个同时兼顾二者的指标，而这就是所谓的 F 值。

用数学术语来说，F 值是精确率和召回率的调和平均值。计算方法如下：

$$F = 2 \cdot \frac{\text{精确率} \cdot \text{召回率}}{\text{精确率} + \text{召回率}} \tag{3}$$

还可以表示为：

$$F = 2 \cdot \frac{C\text{的正确词数}}{C\text{的总词数} + R\text{的总词数}} \tag{4}$$

C 表示候选译文，R 表示参考译文。

下面一起来计算表 4中三个候选译文的 F 值。

表 4: 候选译文 A、B 和 D 的精确率、召回率和 F 值

评测指标	候选译文 A	候选译文 B	候选译文 D
精确率	56%	75%	44%
召回率	63%	75%	50%
F 值	$2 \cdot \frac{56 \cdot 63}{56+63} = 59$	$2 \cdot \frac{75 \cdot 75}{75+75} = 75$	$2 \cdot \frac{44 \cdot 50}{44+50} = 47$

候选译文 A 的 F 值为 59%，候选译文 B 的 F 值为 75%，而候选译文 D 的译文 F 值为 47%。以上结果依然得出，候选译文 B 优于候选译文 A 候选译文 D。

精确率、召回率和 F 值这三个指标的得分越高，则机器翻译的质量越好。然而，这些评测指标只考虑用词，并不考虑词序。

2.5.2 翻译错误率 (TER)

翻译错误率 *(translation error rate, TER)*, 也称作“翻译编辑率”(translation edit rate, TER), 将词序考虑在内。

这种方法基于使用 Levenshtein 距离的单词错误率 (word error rate, WER)。Levenshtein 距离计算不同序列（此处指单词序列）之间的差异, 其定义是“匹配两个序列所需的编辑步骤（包括插入、删除和替换）的最小值”(Koehn 2010: 224)。“单词错误率”根据参考译文的长度将 Levenshtein 距离进行归一化 (Koehn 2010: 225):

$$\text{WER} = \frac{\text{替换词数} + \text{插入词数} + \text{删除词数}}{R\text{的总词数}} \tag{5}$$

R 表示参考译文。

可是, 当词序列或整个从句被移到句子的其他地方, 每个单词的移动都被视为两个错误（即一次删除和一次插入）, 从而导致单词错误率非常高。

为了解决这个问题, “翻译错误率”增加了一次额外操作——移位, 即移动词序列只算一个错误:

$$\text{TER} = \frac{\text{移位词数} + \text{替换词数} + \text{插入词数} + \text{删除词数}}{R\text{的总词数}} \tag{6}$$

R 表示参考译文。

一起回到上述例子。将候选译文 A、B 和 D 与参考译文进行对比 (图 7)。把候选译文 A、B 和 D 改为参考译文, 最少需要多少步?

原文:	Battery pack is attached to the transceiver.
候选译文 A(eTranslation):	Le bloc-batterie est fixé à l’émetteur-récepteur.
候选译文 B(Google Translate):	La batterie est fixée à l’émetteur-récepteur.
候选译文 D(DeepL):	Le bloc-piles est fixé à l’émetteur-récepteur.
参考译文:	La batterie est installée sur l’émetteur-récepteur.

图 7: 原文、候选译文（A、B、D）和参考译文

“翻译错误率”是启发式（或迭代）过程, 这种算法试图通过对比译文的词序列来找到最佳答案（把一个序列转换为另一个序列所需的最少步骤数）。我们可以借助矩阵来手动计算翻译错误率。但为了方便解释, 我们采用更简便但可能不完美的方法[13]: 比较每个候选译文, 并计算匹配、移位、替换、添加和删除的词数。记住, 匹配的词数不计算在内。如前所述, 连字符和撇号被视作单词的“分隔符”。不将这些符号视作分隔符的工具把 l’émetteur-récepteur 当成 1 个单词, 而不是 3 个单词, 因此计算结果也会不一样。

[13]注意, 我们的例子没有移位情况, 因此这里的翻译错误率与单词错误率相等。

2.5.2.1 候选译文 A

表 5: 候选译文 A 变成参考译文所需的操作步骤

操作步骤	被编辑的词	编辑步骤数
匹配	batterie, est, l’, émetteur, récepteur	5
移位		0
替换	le/la, fixé/installée, à/sur	3
插入		0
删除	bloc	1

候选译文 A 的翻译错误率计算如下：

$$\text{TER}_A = \frac{0+3+0+1}{8} = 0.5 = 50\% \tag{7}$$

2.5.2.2 候选译文 B

表 6: 候选译文 B 变成参考译文所需的操作步骤

操作步骤	被编辑的词	编辑步骤数
匹配	la, batterie, est, l’, émetteur, récepteur	6
移位		0
替换	fixée/installée, à/sur	2
插入		0
删除		0

候选译文 B 的翻译错误率计算如下：

$$\text{TER}_B = \frac{0+2+0+0}{8} = 0.75 = 75\% \tag{8}$$

2.5.2.3 候选译文 D

表 7: 候选译文 D 变成参考译文所需的操作步骤

操作步骤	被编辑的词	编辑步骤数
匹配	est, l’, émetteur, récepteur	3
移位		0
替换	le/la, bloc/batterie, fixé/installée, à/sur	4
插入		0
删除	piles	1

候选译文 D 的翻译错误率计算如下：

$$\text{TER}_D = \frac{0+4+0+1}{8} = 0.63 = 63\% \tag{9}$$

候选译文 A 的翻译错误率为 50%，候选译文 B 的翻译错误率为 75%，候选译文 D 的翻译错误率为 63%。因为“单词翻译错误率”和“翻译错误率”都是计算错误所占的百分比，所以这些指标考虑的是不匹配的情况。也就是说，这两者与精确率、召回率和 F 值的解读方式相反，数值越低，则机器翻译的质量越好。因此，由计算结果得出，候选译文 A 为最佳译文。

2.5.3 人工翻译编辑率 (HTER)

有些候选译文虽然与参考译文差异较大，但还是可以接受的，因此以参考译文为准来评测候选译文的方法，可能对机器翻译系统来说严苛且不公正。这时我们可以采用人工翻译编辑率 *(human translation edit rate, HTER)*，即要求评测者对某个候选译文进行译后编辑，然后计算将该候选译文变为译后编辑版本所需的编辑步骤数 (Snover et al. 2006)。

再来看看上述例子。将候选译文 A、B 和 D 与译后编辑版本进行比较，评测者可任选某个候选译文进行编辑 (图 8)：从候选译文 A、B 和 D 到译后编辑版本最少需要多少步？

同理，匹配词数不计算在内。

原文： Battery pack is attached to the transceiver.

候选译文 *A(eTranslation)*： Le bloc-batterie est fixé à l’émetteur-récepteur.

候选译文 *B(Google Translate)*： La batterie est fixée à l’émetteur-récepteur.

候选译文 *D(DeepL)*： Le bloc-piles est fixé à l’émetteur-récepteur.

译后编辑： La batterie est fixée à l’émetteur-récepteur.

图 8: 原文、候选译文（A、B、D）和译后编辑

2.5.3.1 候选译文 A

表 8: 候选译文 A 变成译后编辑版本所需的操作步骤

操作步骤	被编辑的词	编辑步骤数
匹配	batterie, est, à, l’, émetteur, récepteur	6
移位		0
替换	le/la, fixé/fixée	2
插入		0
删除	bloc	1

候选译文 A 的人工翻译编辑率计算如下：

$$\text{HTER}_A = \frac{0+2+0+1}{8} = 0.38 = 38\% \tag{10}$$

2.5.3.2 候选译文 B

表 9: 候选译文 B 变成译后编辑版本所需的操作步骤

操作步骤	被编辑的词	编辑步骤数
匹配	la, batterie, est, fixée, à, l’, émetteur, récepteur	8
移位		0
替换		0
插入		0
删除		0

候选译文 B 的人工翻译编辑率计算如下:

$$\text{HTER}_B = \frac{0+0+0+0}{8} = 0 = 0\% \tag{11}$$

2.5.3.3 候选译文 D

表 10: 候选译文 D 变成译后编辑版本所需的操作步骤

操作步骤	被编辑的词	编辑步骤数
匹配	est, à, l’, émetteur, récepteur	4
移位		0
替换	le/la, bloc/batterie, fixé/fixée	3
插入		0
删除	piles	1

候选译文 D 的人工翻译编辑率计算如下:

$$\text{HTER}_D = \frac{0+3+0+1}{8} = 0.5 = 50\% \tag{12}$$

候选译文 A 的人工翻译编辑率为 38%, 候选译文 B 的人工翻译编辑率为 0%, 候选译文 D 的人工翻译编辑率为为 50%。注意, “翻译错误率” 和 “人工翻译编辑率” 均为错误率, 数值越低, 则机器翻译的质量越好。由计算结果可得, 候选译文 B 的质量最好。

表 11: 候选译文的 TER 值和 HTER 值

评测指标	候选译文 A	候选译文 B	候选译文 D
TER	50%	25%	63%
HTER	38%	0%	50%

现在，比较这个例子 (表 11) 中候选译文的翻译错误率和人工翻译编辑率：所有候选译文的人工翻译编辑率都比翻译错误率值低。这个例子证实了“机器翻译及其译后编辑版本之间的编辑率，大大低于机器翻译和纯人工翻译的参考译文之间的编辑率”(Koehn 2020: 52)。这种差异提醒我们要警惕译后编辑者可能由于时间压力过大，从而导致译后编辑不足的风险。[14]

2.5.4 双语替换评测 (BLEU)

双语替换评测 *(bilingual evaluation understudy, BLEU)* 指候选译文和参考译文共有的 *n*-grams。[15]因此，这种方法同时考虑了匹配词数和词序。其 *n*-grams 的设置范围从 1-gram 到 4-grams，且可以赋予不同权重。

图 9为 René Magritte 的画作《图像的背叛》(*The Treachery of Images*) 的名句的候选译文和参考译文。若以单词 (1-gram) 计算，参考译文含有 6 个单词，其中 5 个出现在候选译文 (若句末的标点符号也算作单词匹配的话)；若以双词 (2-gram) 计算，参考译文含有 5 个双词序列，其中 4 个出现在候选译文 ([is not], [not a], [a pipe], [pipe .])；若以三词 (3-gram) 计算，参考译文含有 4 个三词序列，其中 3 个出现在候选译文 ([is not a], [not a pipe], [a pipe .])；最后，若以四词 (4-gram) 计算，参考译文含有 3 个四词序列，其中 2 个出现在候选译文 ([is not a pipe],[not a pipe.])，第一个四词序列如图 9所示。

原文:	Ceci n’est pas une pipe.
候选译文 (假设):	That is not a pipe .
参考译文:	This is not a pipe .

图 9: 句子 Ceci n’est pas une pipe 的候选译文和参考译文, 示例完全一致的四词序列。

BLEU 值计算的是候选译文与参考译文所匹配的 *n*-grams 数量和候选译文的 *n*-grams 数量的比值，因此该值为精确率。表 12所示为候选译文不同

[14]本章对分享此观点的 Mikel Forcada 致以谢意。
[15]请注意，如前文所述，BLEU 允许使用多个参考译文。

n-gram 的精确率（以比率和小数来表示）。[16]

表 12: 候选译文 That is not a pipe 的精确率 (从 1-grams 到 4-grams) .

评测指标		
精确率 (1-gram)	5/6	0.83
精确率 (2-gram)	4/5	0.80
精确率 (3-gram)	3/4	0.75
精确率 (4-gram)	2/3	0.66

候选译文的总体 BLEU 值便是通过计算不同 *n*-grams 的精确率的几何平均值（一种特殊的“平均值”）而得出，结果略低于 0.76 或 76%。[17]（其实，这一分值非常高，不过例子也很简单。）

我们还应该注意的是，BLEU 值的计算通常基于整个语料库，而非单个句子。有些系统只对有把握的词进行翻译，从而变相提高精确率（就如同上文例子中，害怕犯错而只给出两个回答的学生），因此 BLEU 值还设置了长度惩罚 *(brevity penalty)*，即候选译文和参考译文的词数比（详见Koehn 2020: 227），当候选译文短于参考译文时，该系数就能发挥出作用。不过，图 9的例子没有受到长度惩罚，因为候选译文与参考译文的长度一致。

虽然这一指标通常被称为“BLEU 值”，但是其计算过程涉及很多不同的参数 (Post 2018)，非专业人士难以理解某个自动测评工具如何计算出 BLEU 值。对于想要借助自动评测工具来评估机器译文的译者来说，重要的是，不同译文的评测方法保持一致。简言之，就是确保使用相同的机器翻译评测工具并了解其设置；如果有用户可自定义的设置，则使用该工具时，确保使用相同的自定义设置来对比不同译文。这样，译者就能得到具有可比性的分数。

如表 13所示为图 4候选译文的 BLEU 值，分别由 MutNMT 和 Tilde 的工具计算得出。[18]

由计算结果得出，候选译文 B 的优于候选译文 A。这与截至目前得到的结果一致。不过在实际应用中，不同评测工具算出的 BLEU 值可能出现很大差异，因此用户需要分析导致这种差异的原因。

[16]表 12的表述方式源自 (Koehn 2010: 227)。

[17]读者可采用熟练的电子表格软件来计算几何平均值。专门的 *BLEU* 分数计算器可能还需要考虑其他权重问题，如较长的 *n*-grams，且应该能够对 *n*-grams 精确率为 0 的情况使用平滑方法 (smoothing)。详见Post (2018)。

[18]https://mutnmt.prompsit.com/index；MutNMT 使用 SacreBLEU 算法 (Post 2018)。Tilde 的“交互式 BLEU 值计算器”网址为https://www.letsmt.eu/Bleu.aspx。

表 13: 借助 MutNMT 和 Tilde 计算的候选译文 A、B 和 D 的句级 BLEU 值

BLEU 值计算器	候选译文 A	候选译文 B	候选译文 D
MutNMT	15%	31%	15%
Tilde	50%	61%	47%

2.5.5 ChrF3

字符匹配度 (ChrF) 是基于字符 *n*-grams 的 F 值。因此，该数值以精确率和召回率为基础。回顾一下 F 值的计算公式：

$$F = 2 \cdot \frac{\text{精确率} \cdot \text{召回率}}{\text{精确率} + \text{召回率}} \tag{13}$$

而 ChrF 值的计算公式是：

$$\text{ChrF}\beta = \left(1 + \beta^2\right) \cdot \frac{\text{ChrP} \cdot \text{ChrR}}{\beta^2 \cdot \text{ChrP} + \text{ChrR}} \tag{14}$$

其中，

- ChrP 表示字符 *n*-gram 的精确率，即候选译文的正确字符 *n*-grams 的数量除以其 *n*-grams 的总数，
- ChrR 表示字符 *n*-gram 的召回率，即候选译文的正确字符 *n*-grams 的数量除以参考译文的 *n*-grams 的总数，
- β 是一个权重因子，即表示召回率的影响力是精确度的 β 倍。假设 $\beta = 1$，则召回率和精确率同等重要；ChrF1 是字符 *n*-gram 的精确率和召回率的调和平均数（harmonic mean）(Popović 2015)。[19]

ChrF3 值则是 $\beta = 3$ 时的 ChrF 值的变体，也就是说，召回率的重要性是精确率的三倍。Popović (2015)认为，评测试验表明 ChrF 值，尤其 ChrF3 值，是很有前景的机器翻译自动评测方法。

和 BLEU 值一样，我们不会在此计算 ChrF3 值。不过，可以借助自动评测工具来计算。

表 14展示了通过 MutNMT 计算的候选译文 A、B 和 D 的 ChrF3 值。

由计算结果再次得出，候选译文 B 的优于候选译文 A 和候选译文 C。

[19]注意，这也适用于 F: 到目前为止，所有的 F 值都是 F_1 值，也就是 $\beta = 1$。β 的值可以更改。

表 14: 候选译文 A、B 和 D 的 ChrF3 值

评测指标	候选译文 A	候选译文 B	候选译文 D
ChrF3	64%	69%	49%

2.5.6 提请注意：自动评测指标的结果解读注意事项

在本章节末，我们想提请大家注意一点，那就是确保正确解读自动评测指标的结果。有些结果用小数或百分数来表示（如 0.8 或 80%）。在 0 到 1（或 0% 到 100%）的区间内，对于某个指标（如翻译错误率）来说，0 可能表示最佳，1 可能表示最差；而对于另一个指标（如 BLEU 值），1 可能表示最佳，0 可能表示最差。

此外，在对比不同工具得出的评测指标计算结果时，也应当谨慎，因为看似相同的自动评测指标的算法的不同之处，或许难以被非专业人士察觉。

最后，用户应该知道，若要好好利用各种评测指标，不仅要思考这些结果对其目的的意义，还要懂得如何解读。对比不同的自动评测指标，并将其与人工评测相结合会有所帮助，即使可能存在差异 (Doherty 2017: 134)。在处理非随机顺序呈现的语段时，人工评测对语境的把握会更好，例如，评测者能识别出代词错误，而自动评测指标通常只能基于单词和句子级别来评估选词或句子流畅度。还有一种有意思的方法，是把人工翻译编辑率（即译后编辑的工作量）与译后编辑的工作时长相结合。[20] O'Brien (2025 [本卷]) 简要介绍了如何测量译后编辑工作量。

2.6 类符形符比

和上述指标不同，类符与形符之比 *(type-token ratio, TTR)* 并非用于评测翻译质量，而是全面展现文本的词汇多样性。之所以在此介绍这种方法，是因为已有研究者用该指标对比机器译文与同语言的人工翻译或自然生成的文本，看两者之间的差异程度有多大（参见Toral 2019）。MutNMT 等在线评测工具目前支持使用这一指标。

基本来说,类符形符比用于测量文本或语料库中的词汇多样性(Williamson 2009)。文本的总词数指形符 *(token)* 的数量。但同一个词可能在文本中重复出现，例如，若某个单词在文本中出现三次，则计为 3 个形符，但只能计为 1 个类符 *(type)*。形符与类符之间的数量关系被称为类符形符比，计算方法如下：

$$\text{类符形符比} = \frac{\text{类符数}}{\text{形符数}} \tag{15}$$

[20] 虽然采用两者的平均值说明不了问题，但寻找相关性可能有助于识别最严重的错误。

或

$$\text{类符形符比} = \frac{\text{类符数}}{\text{形符数}} \cdot 100 \tag{16}$$

第一种计算方法的结果范围是 0 到 1，而第二种方法的结果为百分比，范围是 0% 到 100%。类符形符比越高，则文本的词汇就越多样化。

不过，关于类符形符比还要注意以下几点。首先，这种指标对文本的长度非常敏感。其实，文本篇幅越长，限定词和冠词等的重复频率就越高。此外，由于文本都围绕某个主题展开，尤其是专业领域的文本，因此相关术语会重复出现。这样一来，篇幅越长，其类符形符比就越低。所以，为减弱长度对结果的影响，使用者应根据任务将语料划分为固定长度（如 1000 个形符）的片段，分别计算 TTR 再进行标准化。这种标准化的方式可用于对比机器译文语料库与不同长度的同一目标语文本语料库的类符形符比。

其次，虽然在比较同一语言的文本时，词形还原 (lemmatization) 并不重要，但在使用类符形符比来比较两种或两种以上的语言时，必须进行词形还原，因为有些语言的屈折形态比其他语言的更丰富，因此词汇更多样化，而这不过是由于某个动词具有更多形态。不过，如果只是简单使用标准化的类符形符比来比较机器译文和同一语言的其他文本的词汇多样性，那就没有必要进行词形还原。

最后，请记住，类符形符比的数值越大（即词汇更多样化），并不等同于复杂性更高。例如，对比一下 The girl saw a fire.（女孩看见了火。）和 The lexicographer observed the conflagration.（词典编纂者观察到了这场大火。）这两个句子。它们都包含 5 个单词（形符），但前者有 5 个类符，而后者只有 4 个（因为形符 the 出现了 2 次）。因此，第一句的类符形符比为 1（或 100%），而第二句的则为 0.8（或 80%）。可是，虽然第二句的词汇不比第一句的多样化，第二句却更复杂。[21]

如前所述，比较句段级的类符形符比可能没多大意义，但对于文档级或语料库级来说，同一语言文本的标准化类符形符比的比较，则可以为我们提供有用信息，而这也要取决于所处理的文本类型。本章重点关注专业领域翻译。但不同的专业领域需要遵循不同的惯例约定。例如，文学或营销领域翻译的词汇更多样（即类符形符比更高）则质量可能更好，从而提高目标语文本读者的阅读愉悦度；与此相反，技术文本的翻译通常必须遵守某些惯例约定，这往往倾向于降低词汇多样性，以便于用户阅读使用。本章使用的主要例子来自用户手册，也就是说，它应该遵循惯例，如一个术语只表示一个概念，不得有所变化，还应尽量遵循统一的写作范式。例如，若 transceiver 时而被译为 émetteur-récepteur，时而被译为 radio 和 appareil，会导致类符形符比变高，且对最终用户造成困惑。

以上就是希望读者能注意的方面。

[21]作者在此向 Dorothy Kenny 提出的看法致谢。

3 结语

在本章中，我们试图说明从语用维度来评测机器翻译对专业译者或翻译学习者的意义。之所以说语用维度，是因为这种方法不仅将评测当成实现目标的手段，还意味着要根据情况选择不同的方法，通常结合使用人工评测和自动评测。

虽然机器翻译输出的比较贯穿本章内容，但值得注意的是，在实际情况下，专业译者很少有机会选择使用评测指标。相反，他们通常需要迅速判断某个机器翻译解决方案是否适用，或者对机器翻译质量做出总体评价。

因此，我们介绍了如何综合使用人工评测和自动评测来评估机器翻译输出。本章之所以详细介绍了自动评测，是因为我们认为尽管存在局限性，但若理解得当并结合人工评测，自动评测就可以得到有效利用。

References

Aziz, Wilker, Sheila Castilho & Lucia Specia. 2012. PET: A tool for post-editing and assessing machine translation. In *Proceedings of the eight international conference on language resources and evaluation (LREC'12)*, 3982–3987.

Castilho, Sheila. 2020. On the same page? Comparing inter-annotator agreement in sentence and document level human machine translation evaluation. In *Proceedings of the 5th conference on machine translation (WMT)*, 1150–1159. https://aclanthology.org/2020.wmt-1.137.pdf.

Castilho, Sheila, Stephen Doherty, Federico Gaspari & Joss Moorkens. 2018. Approaches to human and machine translation quality assessment. In Federico Gaspari Joss Moorkens Sheila Castilho & Stephen Doherty (eds.), *Translation quality assessment: From principles to practice*, 9–38. Cham: Springer.

Castilho, Sheila, Joss Moorkens, Federico Gaspari, Iacer Calixto, John Tinsley & Andy Way. 2017. Is neural machine translation the new state of the art? *The Prague Bulletin of Mathematical Linguistics* 108. 109–120. DOI: 10.1515/pralin-2017-0013.

Castilho, Sheila & Sharon O'Brien. 2016. Evaluating the impact of light post-editing on usability. In *10th international conference on language resources and evaluation (LREC)*, 310–316. May 2016, Portorož, Slovenia. ELRA.

Doherty, Stephen. 2017. Issues in human and automatic translation quality assessment. In Dorothy Kenny (ed.), *Human issues in translation technology*, 131–148. London: Routledge.

Drugan, Joanna. 2013. *Quality in professional translation: Assessment and improvement*. London: Bloomsbury.

Gouadec, Daniel. 2010. Quality in translation. In *Handbook of translation studies. Volume 1*, 270–275. John Benjamins Publishing Company.

Grbić, Nadja. 2008. Constructing interpreting quality. *Interpreting* 10(2). 232–257.

House, Juliane. 2015. *Translation quality assessment: Past and present*. London: Routledge.

Kenny, Dorothy. 2025. 人工翻译和机器翻译. In Dorothy Kenny (ed.), 机器翻译知识普及: 为人工智能时代的用户赋能, 19–38. Berlin: Language Science Press. DOI: 10.5281/zenodo.14922287.

Koehn, Philipp. 2010. *Statistical Machine Translation*. Cambridge: Cambridge University Press.

Koehn, Philipp. 2020. *Neural Machine Translation*. Cambridge: Cambridge University Press.

Mariana, Valerie, Troy Cox & Alan Melby. 2015. The multidimensional quality metric (MQM) framework: A new framework for translation quality assessment. *The Journal of Specialised Translation* 23. 137–161.

Moorkens, Joss. 2018. What to expect from neural machine translation: A practical in-class translation evaluation exercise. *The Interpreter and Translator Trainer* 12(4). 375–387.

Moorkens, Joss. 2025. 伦理道德与机器翻译. In Dorothy Kenny (ed.), 机器翻译知识普及: 为人工智能时代的用户赋能, 95–110. Berlin: Language Science Press. DOI: 10.5281/zenodo.14922295.

O'Brien, Sharon. 2025. 如何处理机器翻译的错误：译后编辑. In Dorothy Kenny (ed.), 机器翻译知识普及: 为人工智能时代的用户赋能, 83–94. Berlin: Language Science Press. DOI: 10.5281/zenodo.14922293.

Popović, Maja. 2015. Chrf: Character n-gram f-score for automatic MT evaluation. In *Proceedings of the tenth workshop on statistical machine translation*, 392–395. Association for Computational Linguistics. 10.18653/v1/W15-3049.

Post, Matt. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the third conference on machine translation (WMT), Volume 1: Research papers*, 186–191. Association for Computational Linguistics. DOI: 10.18653/v1/W18-6319.

Qin, Ying & Lucia Specia. 2015. Truly exploring multiple references for machine translation evaluation. In *Proceedings of the 18th annual conference of the European Association for Machine Translation*, 113–120. https://aclanthology.org/W15-4915/.

Snover, Matthew, Bonnie Dorr, Rich Schwartz, Linnea Micciulla & John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th conference of the Association for Machine Translation in the Americas: Technical papers*, 223–231. Cambridge, Massachusetts: Association

for Machine Translation in the Americas. https://aclanthology.org/2006.amta-papers.25/.

Toral, Antonio. 2019. Post-editese: An exacerbated translationese. In *Proceedings of machine translation summit XVII*, 273–281. EAMT. https://www.aclweb.org/anthology/W19-6627/.

Williamson, Graham. 2009. *Type-token ratio*. Last retrieved 5 Dec. 2020. https://www.sltinfo.com/wp-content/uploads/2014/01/type-token-ratio.pdf.

Yamada, Masaru. 2019. The impact of Google neural machine translation on post-editing by student translators. *The Journal of Specialised Translation* 31(2019). 87–106.