神经机器翻译用于科普特语-法语：低资源古代语言的策略

Nasma Chaoui and Richard Khoury
Department of Computer Science and Software Engineering
Université Laval
Québec City, Canada
nasma.chaoui.1@ulaval.ca, richard.khoury@ift.ulaval.ca

Abstract

本文首次系统地研究了从科普特语翻译到法语的策略。我们的综合流程系统地评估了：枢纽翻译与直接翻译、预训练的影响、多版本微调的好处以及模型对噪声的鲁棒性。利用对齐的圣经语料库，我们证明了使用风格多样且考虑噪声的训练语料库进行微调能够显著提升翻译质量。我们的研究结果为开发用于历史语言的翻译工具提供了重要的实用见解。

1 介绍

神经机器翻译（NMT）凭借其稳步增长，对计算语言学领域产生了深远影响。变压器模型和大规模多语言系统使得数百种语言对的翻译成为可能。然而，许多历史语言——尤其是数据稀缺的古代语言——尚未从这些进步中受益。科普特语便是其中一种。作为埃及语言家族的最后阶段，理解和翻译它对于推进埃及学和早期基督教研究等领域至关重要。

因此，该语言在世界各地的大学中被研究；根据一项统计 [1]，13 个国家的 50 多所大学提供科普特语课程。但尽管有这样的国际兴趣，目前针对科普特语的神经机器翻译工具很少，而且仅限于翻译成英语或阿拉伯语。

本文旨在通过研究一系列的神经机器翻译（NMT）策略来实现从科普特语到法语的翻译，以填补这一空白。在这个过程中，我们将重点关注四个重要问题：（1）我们应该直接翻译到目标语言还是使用英语作为中介语言；（2）我们应该如何选择语言模型来针对这个任务进行微调；（3）我们是否应该为模型提供一个多样的译文集合用于一个段落；以及（4）我们如何使模型对原稿中的噪声具有鲁棒性。虽然我们只专注于科普特语到法语的 NMT，这些问题具有普遍的兴趣，我们的研究结果将有助于指导任何古代-现代语言对之间的 NMT 工具的发展。

本文的其余部分组织如下：第 2 节回顾了相关的科普特语自然语言处理工作。第 3 节详细阐述了我们计划用来探讨四个问题的方法。第 4 节展示了我们的实验设置：我们的数据集、NMT 模型和评估指标，第 5 节展示了我们在关于四个问题的实验中获得的结果。最后，第 6 节给出了一些结论性评论。我们的数据集和代码可以在我们的 GitHub 账户 [2] 上获得，而我们训练的最佳科普特语 → 法语 NMT 模型则在 HuggingFace [3] [4] 上分发。

2 背景

科普特语占据了一个独特的位置，是埃及语系的最后阶段，从古埃及语到世俗体文字桥接了数千年的历史。对于埃及学、圣经研究和早期修道主义研究，它具有重要的历史和文化意义，因为许多科普特手稿阐明了这些领域。

近年来，在建立科普特语的数字资源方面取得了显著进展。科普特语 SCRIPTORIUM 项目 (Schroeder and Zeldes, 2016) 作为一项基础性工作而突出，为多个语言层次提供了带注释的语料库，包括词性标注、词形还原、形态分析和句法解析 (Zeldes and Schroeder, 2016) 。同样地，诸如 BabyLemmatizer 这样的工具已经扩展了科普特语、古埃及民用书写体和早期埃及语的词形还原和词性标注能力，显示了埃及语族内部共同的语言复杂性 (Sahala and Lincke, 2025) 。

最近的工作也在埃及语言家族中探索了神经机器翻译，特别是聚焦于象形文字文本。值得注意的是，(Cao et al., 2024) 通过调整多语言 M2M-100 模型开发了一种象形文字 Transformer，将古埃及象形文字翻译成德语和英语。他们的方法涉及从《埃及语言词典》中构建一个精心整理的数据集，其中包括象形文字铭文与音译、词性标记和翻译配对。他们依靠音译

[1] 由 https://www.cmcl.it/~iacs/course.htm 汇编的清单。

[2] https://github.com/chaouin/coptic-french-nmt

[3] https://huggingface.co/chaouin/coptic-french-translation-helsinki

[4] https://huggingface.co/chaouin/coptic-french-translation-hiero

（罗马化）来使象形文字序列与现代自然语言处理模型词汇表兼容，并展示了从现有的多语言架构进行迁移学习可以在低资源古代文献上取得可观的结果。

(Saeed et al., 2024) 的作者引入了迄今为止最大的科普特文平行语料库，包含约 24,000 对科普特语-英语句子对和 8,000 对科普特语-阿拉伯语句子对，并利用它训练了第一个从科普特语到阿拉伯语和英语的神经机器翻译（NMT）系统。他们使用了基本的序列到序列模型以及微调的多语言模型，例如 mT5 和 M2M-100。他们的工作强调了由于科普特字符缺乏标准词汇表而进行拼音化的必要性。类似地，(Enis and Megalaa, 2024) 通过对来自 Coptic SCRIPTORIUM 的标准化和拼音化文本微调多语言模型，实施了科普特语到英语的 NMT。他们采用了反向翻译并探索了多种数据表示，展示了形态学标准化和标记化对处理科普特文黏着结构的关键影响。

3 方法论

在实现我们的科普特 → 法语 NMT 模型时，我们力求研究四个基本设计问题，这些问题在大多数古代到现代语言翻译项目中很常见。因此，我们的方法分为四个阶段，每个阶段都基于前一个阶段的发现。

首先，我们想要回答一个基本问题：对科普特语 → 法语翻译的神经机器翻译（NMT）模型进行微调是否值得？毕竟，虽然没有这样的模型存在，但有其他替代方案。第一个是使用一个更通用的多语言 NMT 模型，它可以在没有任何微调的情况下直接使用。另一种方案是使用一个可以将古代语言翻译成另一种枢纽语言的 NMT 模型，并提示它用我们的目标现代语言回答。在我们的情况下，这意味着使用一个现成的科普特语 → 英语 NMT 模型，并提示它用法语回答。最后，由于今天训练于现代语言对之间的 NMT 模型层出不穷，我们可以再次从一个可以将古代语言翻译成枢纽语言的 NMT 模型开始，并使用第二个 NMT 模型将内容翻译成我们的目标现代语言。在我们的案例中，我们可以串联一个科普特语 → 英语和一个英语 → 法语 NMT 模型来实现这一点。因此，我们研究的第一阶段是比较微调的科普特语 → 法语 NMT 模型和这三种替代方案所产生的翻译，以观察这四个选择之间的差异。

虽然古代现代神经机器翻译 (NMT) 模型并不多，但有几个可用，因此出现了一个问题：如何选择其中一个进行微调？在这个决策中存在技术和语言的考虑。从技术角度来看，不同的可用模型可能有不同数量的参数，并且使用不同数量的数据进行训练（但鉴于古代语言数据的稀缺，它们通常不会有不同的数据来源）。从语言上讲，当没有模型存在于我们的古代和现代语言对之间时，有三个选项可用。第一个是采用一个训练用以将古代语言翻译成不同现代语言的模型，并将其微调到我们的目标语言，或者在我们的情况下，微调一个科普特语到英语的模型以翻译成法语。第二个选项是采用一个训练用以翻译到现代语言的模型，并将其微调到古代语言。在我们的情况下，这意味着采用任何可用的多语言 NMT 模型，它们训练用于法语，并微调以自科普特语翻译。最后，可能存在针对同一语言家族中不同古代语言的 NMT 翻译模型，为我们提供第三个选项。如前所述，科普特语是埃及语言的最后阶段，因此一个为该语言的早期阶段训练的 NMT 模型可以被微调用于科普特语。

如果我们想对现有模型进行微调，我们需要一个古代语言和现代语言的平行文本语料库。然而，由于各种原因，古代语言文本可能有多种不同方式被翻译成现代语言。因此，我们进行了一项实验，以确定微调是否可以从这种风格多样性中受益。我们的假设是，对于同一原文有多个不同的翻译，这将使得能够从稀缺的古代语言资源中构建更大的训练语料库，同时也使得神经机器翻译模型能够更好地学习句法变化和词汇同义的概念。

3.1 训练对噪声的鲁棒性

古代语言文本由于时间的流逝往往会遭受损坏。在我们最后的实验中，我们考虑使用噪声数据训练 NMT 模型作为应对噪声的鲁棒性构建的一种方法。我们考虑两种噪声来源。第一种是由于文本的碎片缺失或损坏导致的缺漏 (Levine et al., 2024) 。我们可以通过随机用“缺失字符”符号替换文本中的字符来在我们的训练数据中引入这个问题。我们研究的第二个噪声来源是字母替换。这一问题来自两个不同的来源：一方面，某些语言中的字符可能看起来相似，手写可能会导致歧义，另一方面，用于将手稿数字化的光学字符识别（OCR）系统容易识别错误字符，有研究发现古代文本 OCR 中的字符识别准确率可能低至 62 % (Tzogka et al., 2021) 。我们可以通过创建一组看起来相似的字符列表，然后随机用列表中相似的字符替换一个字符来模拟这个问题。最后一种错误来源是拼写变化和错误，我们通过随机切换两个相邻字符的顺序来模拟这一问题。通过控制三种随机替换的比例，我们可以进一步实验不同比例的噪声对数据的影响。

4 实验设置

4.1 数据集和预处理

为了构建我们的科普特语-法语平行语料库，我们首先从 Coptic SCRIPTORIUM 项目 (Schroeder and Zeldes, 2016) 中提取了科普特文本，该项目提供了丰富注释的 PAULA XML 文件。我们专注于最完整和标准化的科普特语料库–Shahidic 新旧约圣经。我们自动解析了 XML 文件以提取诗句级别的科普特文本及其诗句 ID（圣经书籍、章节和诗句编号）。

由于大多数预训练的 NMT 模型与原始的科普特语 Unicode 脚本不兼容，我们使用 `uroman` 工具 [5] 将科普特语数据进行了罗马化，这是由 (Amrhein and Sennrich, 2020) 建议并由 (Enis and Megalaa, 2024) 用于科普特语的方法。这确保了与现有 Transformer 模型的分词器和词汇的兼容性。

对于法文文本，我们收集了三个公共领域的圣经译本：路易 · 塞贡德版、克拉蓬版和达比版。选择这些译本是因为它们在风格上的多样性、现代相关性和免费可获得性。与科普特文本文的对齐是通过使用章节 ID 自动进行的。

接下来，我们进行了一个清理步骤，以确保高质量的平行语料库。我们删除了任何缺少科普特文或法文翻译的诗句对。我们还通过筛选出那些科普特文段落只由省略号（例如，“[...]”）组成的条目来移除结构上不匹配的对，这表明诗句不完整或缺失。另外，我们丢弃了那些法文翻译缺失或为空白的诗句，并从法文诗句中去除开头的内联注释（例如，（1.2）），以仅保留规范内容。所有被删除的条目都单独记录以确保结果的可重复性。清理之后，我们保留了 23,561 对已经对齐的 “Segond” 和 “Darby” 版本的诗句，以及科兰蓬版本的 23,718 对，共从 63 本圣经书中得到 70,840 对对齐的诗句。

我们将这个数据集拆分为训练和测试语料库。我们选择了按书籍级别拆分，保留哥林多前书、马可福音、加拉太书和希伯来书作为测试数据，其余 59 本书作为训练数据。这总共代表了 1,460 节科普特语测试经文，每节都有 3 个法语翻译。此设置防止数据污染，即来自同一本书的相似经文可能会在训练和测试语料库中找到。它还模拟了真实世界的应用场景，即发现了一个从未出现过的新科普特文献。

我们选择了四个预训练的 NMT 模型进行研究。它们与我们的科普特 → 法语翻译挑战最相关。

除非另有说明，所有模型在我们的训练语料库上进行了 15 个周期的微调。

[5] https://github.com/isi-nlp/uroman

4.2 指标

正如我们的数据集明确展示的那样，一句话科普特语声明可以有几种不同但等效的法语翻译。因此，一个好的模型应能够正确翻译原声明的意图，即使其措辞与标准参考不一致。为评估翻译质量，我们侧重于衡量语义相似性而非词汇重合度的指标。

- BERTScore (Zhang et al., 2020) 使用预训练的 BERT 模型的上下文嵌入计算词级相似性。得分范围从 0 到 1，其中较高的值表示候选句子和参考句子之间的语义相似性更大。它非常适合评估语义等价性，特别是在存在多个有效版本的陈述时。

- BLEURT (Sellam et al., 2020) 是一种学习的评估指标，它微调基于 BERT 的模型来预测人类对文本质量的判断。它将语义相似性、流畅性和语法性结合成一个单一的分数，并已被广泛用于评估机器翻译输出 (Pu et al., 2021) 。BLEURT 分数没有固定范围，但通常集中在 [0, 1] 之间，且数值越高表示质量越好。

- COMET (Rei et al., 2020) 是一个专门为机器翻译设计的学习评价指标。它训练一个多语言的 BERT 模型来预测人类对翻译质量的判断。COMET 分数的范围是 0 到 1，分数越高，表明对翻译的人的欣赏程度越高。

- METEOR (Banerjee and Lavie, 2005) 是一个基于单词匹配的机器翻译评估指标，但它考虑了同义词转换、词干提取和释义，并进行了调整，以奖励语义上有效的表达方式，而不是精确的词汇匹配。METEOR 分数在 0 和 1 之间标准化，1 表示完全匹配。

通过综合考虑这些指标，我们可以获得更全面的翻译质量评估，该评估不仅考虑对意义的保留和文本质量，并且对语义等价的词汇变化具有容忍度。

5 结果与分析

5.1 基线

我们首先使用第 **??** 节的模型建立一组基线。为此，每个模型通过翻译我们测试语料库中的书籍来评估，用他们被训练的最相关语言对进行：Coptic → 英语用于 Enis，Coptic → 法语用于 Helsinki，和英语 → 法语用于 T5。由于不存在圣经文本的象形文字形式来评估 Hiero 模型，我们使用其自身的测试语料库句子对其进

行了测试，该语料库由 50 对象形文字 → 英语对和 50 对象形文字 → 德语对组成。

Model	BERTScore	BLEURT	COMET	METEOR
Enis	0.632	0.344	0.328	0.119
Helsinki	0.595	0.136	0.248	0.082
T5	0.809	0.501	0.673	0.429
Hiero	0.833	0.555	0.664	0.496

Table 1: 基线预训练模型性能。

5.2 微调的相关性

我们的第一个实验比较了实现科普特语 → 法语翻译的四种不同策略，这些策略在第 **??** 节中有所描述。首先是微调现有的科普特语翻译模型 Enis 模型以实现法语目标。第二种策略是使用包含科普特语和法语的多语言模型，即 Helsinki 模型。第三种策略是不经过微调，以零样本方式提示 Enis 模型翻译为法语。最后，我们将使用英语作为中介语言，通过使用 Enis 将科普特语翻译成英语，然后使用 T5 将所得输出翻译成法语。

Approach	BERTScore	BLEURT	COMET	METEOR
Fine-tuning	0.820	0.402	0.613	0.467
Multilingual	0.595	0.136	0.248	0.082
Prompting	0.606	0.149	0.238	0.041
Pivot language	0.620	0.115	0.274	0.100

Table 2: 翻译策略的比较

表格 2 中的结果清楚地表明，与其他三种方法相比，微调的好处证明了额外工作的合理性。经过额外训练后，Enis 模型在法语翻译中比在我们的基线结果表格 1 中对英语的翻译生成了更好的结果。由于解码期望的不匹配，在不进行微调的情况下提示 Enis 翻译为法语，结果是在三次 Enis 测试中最差的。微调为一对语言的模型也优于多语言模型，这一结果与文献一致，文献发现多语言模型在低资源语言（如科普特语）上的表现次优 (Joshi et al., 2025) 。最后，有趣的是，Enis-T5 组合在所有指标上的表现都比 Enis 或 T5 在表格 1 中的基线测试中要差。这表明每一步中的翻译错误由于翻译序列而累积，降低了整体准确性。这与 (Paul et al., 2013) 的观察一致。

5.3 预训练模型选择

已经证明为科普特语 → 法语微调模型是值得的，我们的第二个实验探索了第 **??** 节的问题，即应该微调哪个模型。对于这个实验，我们在整个训练语料库上微调每个模型。这使得 Enis 可以学习翻译成法语，T5 可以从科普特语翻译，Hiero 可以从科普特语翻译成法语。尽管 Helsinki 已经在其 100 种语言中包括科普特语和法语的训练，但它也被微调以专注于这两种语言。

Model	BERTScore	BLEURT	COMET	METEOR
Enis	0.820	0.402	0.613	0.467
Helsinki	0.850	0.564	0.731	0.563
T5	0.763	0.240	0.474	0.325
Hiero	0.832	0.498	0.681	0.501

Table 3: 微调模型的比较。

表 3 展示了我们的测试语料翻译结果，显示 Helsinki 在所有四个指标上都优于其他模型。这种性能可归因于两个关键优势。首先，其预训练语料包括源语言和目标语言，而其他模型缺乏其中一种或两种语言，必须从头开始学习。其次，它是预训练语料最接近我们任务的模型，也同样是圣经语料。Helsinki 模型相对较大，拥有 248M 参数，可能有助于其出色的性能。然而，我们不能排除它可能对翻译圣经文本产生过拟合的可能性；由于我们没有非圣经的科普特-法语翻译进行测试，无法检验这种可能性。

在所有指标上表现第二好的模型是 Hiero。有趣的是，这个模型与 Helsinki 的属性相反：它既没有接受科普特语、法语，也没有圣经文本的训练。然而，它接受过象形文字的训练，而象形文字是一种与科普特语相关的语言。这似乎表明在一种语言上学到的知识能够很好地传递到相关语言上。在使用中介语言的 NMT 案例中，(Babych et al., 2007) 也曾指出类似的观察。Enis 模型位居第三。虽然该模型受益于预训练时使用了科普特语，但它比 Hiero 模型要小，Enis 只有 77.1M 参数，而 Hiero 有 484M 参数，因此在学习新语言方面不如 Hiero。最后，T5 模型拥有 223M 参数，在其预训练数据中既没有科普特语也没有相关语言，因此在微调结果中表现最差。

5.4 微调数据集选择

为了验证我们在第 **??** 节中的假设，我们对第 5.3 节中的两个最佳模型，即 Helsinki 和 Hiero，分别在我们的三个法语翻译文本上进行了微调。对于这个实验，由于每个单一来源模型训练的大约是完整训练数据集的三分之一，我们将这些模型训练了 45 个周期，而不是对整个数据集训练时的 15 个周期。我们分别在测试语料库的每个译本上测试了每个模型，以观察不同训练产生的影响。

表 4 和 5 中给出的结果显示，模型在特定数据集翻译上训练后，在进行相同翻译测试时，在我们的四个指标之一——BERTScore 和 METEOR 中表现最好。这两个指标重视与正

Test set	Training set	BERTScore	BLEURT	COMET	METEOR
Crampon	All datasets	0.836	0.549	0.723	0.513
Crampon	Crampon only	0.849	0.537	0.715	0.583
Crampon	Darby only	0.821	0.494	0.690	0.460
Crampon	Segond only	0.825	0.519	0.703	0.472
Darby	All datasets	0.858	0.576	0.737	0.587
Darby	Crampon only	0.825	0.511	0.696	0.493
Darby	Darby only	0.868	0.558	0.729	0.615
Darby	Segond only	0.846	0.535	0.710	0.546
Segond	All datasets	0.856	0.569	0.732	0.587
Segond	Crampon only	0.828	0.517	0.699	0.514
Segond	Darby only	0.844	0.515	0.701	0.550
Segond	Segond only	0.865	0.564	0.727	0.611

Table 4: 赫尔辛基数据集选择比较。

Test set	Training set	BERTScore	BLEURT	COMET	METEOR
Crampon	All datasets	0.815	0.480	0.672	0.445
Crampon	Crampon only	0.825	0.458	0.655	0.493
Crampon	Darby only	0.805	0.433	0.642	0.410
Crampon	Segond only	0.806	0.448	0.646	0.413
Darby	All datasets	0.844	0.517	0.692	0.535
Darby	Crampon only	0.812	0.450	0.648	0.444
Darby	Darby only	0.847	0.493	0.678	0.547
Darby	Segond only	0.830	0.472	0.660	0.489
Segond	All datasets	0.837	0.495	0.680	0.519
Segond	Crampon only	0.810	0.442	0.642	0.443
Segond	Darby only	0.826	0.451	0.648	0.487
Segond	Segond only	0.839	0.479	0.663	0.521

Table 5: 选择数据集对比在 Hiero 上。

确翻译的语义相似性。通过在相同翻译上进行训练和测试，模型学习到正确的措辞选择和翻译决策，因此能够生成与预期非常相似的翻译。相比之下，在一个翻译上训练而在另一个翻译上测试时，所有这些系统的 BERTScore 和 METEOR 表现都要弱得多，这表明它们生成的翻译与预期结果的措辞不同。

另一方面，无论测试于哪个数据集翻译，针对所有三个数据集翻译进行训练的模型总是能获得最高的 BLEURT 和 COMET 分数。这两个指标经过调整以模拟人类对生成文本质量的认可。通过针对每个科普特语句的多种可能翻译进行训练，该模型学会了总体上生成更好的翻译。此外，该模型始终获得第二高的 BERTScore 和 METEOR 分数，仅次于在相同测试语料库翻译上训练的模型，但领先于在其他翻译上训练的两个模型。因此，尽管该模型生成的措辞相较于正确翻译而言并非最佳，但总是比任何其他非专门生成该特定翻译的模型更接近。总体而言，这表明训练数据中的风格多样性在推理时提高了生成质量，尤其是在多种有效翻译都可接受的情况下。

5.5 训练对噪声的鲁棒性

如第 3.1 节所述，我们通过在训练和测试数据中引入噪声来模拟手稿的真实退化。为了模拟文本损坏，科普特文诗句中的每个字符会经过三个独立的噪声处理：2 % 的删除概率（模拟残缺）、2 % 的与相邻字符交换的概率（模拟拼写错误）、以及 10 % 的被视觉上相似的字符替换的概率（模拟 OCR 错误），使用图 1 的视觉混淆映射。我们不改变诗句中的法语翻译。通过创建 5 个版本的语料库，我们尝试不同程度的模拟退化，设置诗句被损坏的概率为 0 % 、10 % 、30 % 、50 % 和 100 % 。

ⲁ → ⲟ, ⲉ　ⲟ → ⲁ, ⲉ　ⲉ → ⲁ, ⲟ　ⲓ → ⲏ, ⲩ

ⲏ → ⲓ, ⲉ　ⲩ → ⲓ, ⲛ　ⲥ → ⲓ, ⲏ　ⲫ → ⲕ, ⲭ

ⲕ → ⲫ, ⲭ　ⲭ → ⲫ, ⲕ　ⲧ → ⲧ, ⲑ　ⲧ → ⲧ, ⲑ

ⲑ → ⲧ, ⲧ　ⲛ → ⲩ　ⲙ → ⲛ　ⲣ → ⲩ, ⲙ

Figure 1: 科普特字符的视觉混淆。

Test noise	Train noise	BERTScore	BLEURT	COMET	METEOR
0 %	0 %	0.850	0.564	0.731	0.563
0 %	10 %	0.852	0.568	0.733	0.564
0 %	30 %	0.849	0.560	0.727	0.560
0 %	50 %	0.848	0.553	0.723	0.559
0 %	100 %	0.846	0.547	0.719	0.546
10 %	0 %	0.844	0.545	0.717	0.549
10 %	10 %	0.849	0.557	0.725	0.557
10 %	30 %	0.847	0.553	0.722	0.555
10 %	50 %	0.846	0.545	0.718	0.554
10 %	100 %	0.844	0.544	0.717	0.543
30 %	0 %	0.829	0.492	0.679	0.511
30 %	10 %	0.842	0.534	0.708	0.538
30 %	30 %	0.842	0.539	0.711	0.545
30 %	50 %	0.843	0.536	0.712	0.546
30 %	100 %	0.841	0.533	0.709	0.536
50 %	0 %	0.817	0.457	0.650	0.478
50 %	10 %	0.837	0.514	0.695	0.522
50 %	30 %	0.838	0.522	0.701	0.533
50 %	50 %	0.838	0.519	0.698	0.533
50 %	100 %	0.838	0.524	0.702	0.529
100 %	0 %	0.788	0.366	0.579	0.411
100 %	10 %	0.822	0.471	0.665	0.488
100 %	30 %	0.826	0.491	0.675	0.502
100 %	50 %	0.829	0.496	0.683	0.509
100 %	100 %	0.831	0.505	0.688	0.513

Table 6: 关于赫尔辛基对噪声的鲁棒性比较。

表 6 和 7 中的结果显示了几个明显的趋势。首先，以任何噪声水平训练的两个模型在干净的测试数据上表现更好，并且随着噪声的增加，它们的得分逐渐下降。这表明带有错误或缺失的文本相比干净的文本更难翻译。同样，经过低噪声训练的 Helsinki 和 Hiero 模型在低噪声数据上的表现更好，而经过更多噪声训练的模型在噪声数据上的表现更佳。对于 Helsinki，早期由 10 % 噪声训练的模型领先，然后 50 % 模型在中等噪声环境中占据优势，而 100 % 模型在最后的测试中表现最佳。对于 Hiero，0 % 模型在早期表现最佳，30 % 模型在中期领先，100 % 模型在最后的测试中占据主导地位。

真正发生的情况是，用较少噪声训练的模型在处理干净数据时表现更好，但它们也更脆弱：随着噪声增加，它们的得分急剧下降。表 8 显示了每个模型和指标在 0 % 和 100 % 测试噪声之间的相对得分下降。用 0 % 和 10 % 噪声训练的模型平均损失大约为 20 %——Hiero

Test noise	Train noise	BERTScore	BLEURT	COMET	METEOR
0 %	0 %	0.832	0.497	0.682	0.500
0 %	10 %	0.832	0.496	0.679	0.496
0 %	30 %	0.831	0.498	0.679	0.494
0 %	50 %	0.830	0.492	0.674	0.492
0 %	100 %	0.827	0.482	0.668	0.482
10 %	0 %	0.825	0.474	0.665	0.485
10 %	10 %	0.828	0.484	0.670	0.488
10 %	30 %	0.829	0.489	0.674	0.489
10 %	50 %	0.829	0.485	0.671	0.488
10 %	100 %	0.826	0.478	0.665	0.480
30 %	0 %	0.809	0.412	0.619	0.445
30 %	10 %	0.822	0.452	0.649	0.470
30 %	30 %	0.825	0.465	0.658	0.476
30 %	50 %	0.824	0.466	0.656	0.476
30 %	100 %	0.824	0.466	0.657	0.473
50 %	0 %	0.795	0.364	0.582	0.416
50 %	10 %	0.815	0.426	0.630	0.453
50 %	30 %	0.819	0.447	0.643	0.463
50 %	50 %	0.821	0.452	0.647	0.466
50 %	100 %	0.821	0.456	0.650	0.466
100 %	0 %	0.762	0.232	0.486	0.335
100 %	10 %	0.800	0.366	0.592	0.411
100 %	30 %	0.810	0.405	0.616	0.436
100 %	50 %	0.813	0.419	0.627	0.445
100 %	100 %	0.815	0.426	0.629	0.451

Table 7: 对噪声的鲁棒性比较用于 Hiero。

的 BLEURT 得分甚至下降了 53 %。相比之下，用 50 % 或 100 % 噪声训练的模型下降不到 15 %，平均仅 6 %。这证实，受控噪声注入可以创建对现实世界退化（如 OCR 错误或片段化手稿）更具弹性的 NMT 模型，但代价是对干净手稿的性能较低。因此，最优选择取决于被翻译手稿的特征，但一个良好的通用折中值似乎是在 50 % 训练噪声。

Model	Training noise	BERTScore	BLEURT	COMET	METEOR
Helsinki	0 %	7.3 %	35.1 %	20.8 %	27.0 %
Helsinki	10 %	3.5 %	17.1 %	9.3 %	13.5 %
Helsinki	30 %	2.7 %	12.3 %	7.2 %	10.4 %
Helsinki	50 %	2.2 %	10.3 %	5.5 %	8.9 %
Helsinki	100 %	1.8 %	7.7 %	4.3 %	6.0 %
Hiero	0 %	8.4 %	53.3 %	28.7 %	33.0 %
Hiero	10 %	3.8 %	26.2 %	12.8 %	17.1 %
Hiero	30 %	2.5 %	18.7 %	9.3 %	11.7 %
Hiero	50 %	2.0 %	14.8 %	7.0 %	9.6 %
Hiero	100 %	2 %	11.6 %	5.8 %	6.4 %

Table 8: 性能下降从 0 % 到 100 % 噪声。

6 结论

本文首次系统研究了用于科普特语和法语之间的神经机器翻译。通过四个实验，我们概述了为这个新语对构建翻译模型的最佳实践。首先，我们展示了相比使用预训练的多语言模型或通过中间语言翻译，特定任务的微调显著提高了翻译质量。接下来，我们发现对已经在古代语言或同一语系（如象形文字）中相关语言上训练过的模型进行微调能获得更好的结果。第三，我们考察了对同一文本进行多次翻译的效果，发现全部包含这些翻译增强了翻译质量。最后，我们测试了数据集中噪声的影响，发现最佳的翻译质量与稳健性之间的权衡是在使用 50 % 受损的诗节进行训练时实现的。这导致了两个强大的 NMT 模型的发布，一个基于 (Tiedemann et al., 2023) 的多语言模型，另一个基于 (Cao et al., 2024) 的象形文字翻译器，二者均在包含 50 % 噪声的三个法语版本的训练语料库上进行了微调。总之，我们的研究结果超越了科普特语到法语的翻译，提供了在其他低资源或古代语言环境中构建稳健的 NMT 系统的实用指南。

7

致谢 这项研究得以进行是由于加拿大自然科学与工程研究委员会（NSERC）的资助。

References

Chantal Amrhein and Rico Sennrich. 2020. On Romanization for model transfer between scripts in neural machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2461–2469, Online. ACL.

Bogdan Babych, Anthony Hartley, and Serge Sharoff. 2007. Translating from under-resourced languages: comparing direct transfer against pivot translation. In *Proceedings of Machine Translation Summit XI: Papers*, Copenhagen, Denmark.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. ACL.

Mattia Cao, Nicola De Cao, Angelo Colonna, and Alessandro Lenci. 2024. Deep learning meets egyptology: a hieroglyphic transformer for translating ancient egyptian. In *Proceedings of the 1st Workshop on Machine Learning for Ancient Languages (ML4AL 2024)*, pages 71–86.

Maxim Enis and Andrew Megalaa. 2024. Ancient voices, modern technology: Low-resource neural machine translation for coptic texts. *Coptic Translator*, pages 1–15.

Raviraj Joshi, Kanishk Singla, Anusha Kamath, Raunak Kalani, Rakesh Paul, Utkarsh Vaidya, Sanjay Singh Chauhan, Niranjan Wartikar, and Eileen Long. 2025. Adapting multilingual LLMs to low-resource languages using continued pre-training and synthetic corpus: A case study for Hindi LLMs. In *Proceedings of the First Workshop on Natural Language Processing for Indo-Aryan and Dravidian Languages*, pages 50–57, Abu Dhabi. ACL.

Lauren Levine, Cindy Li, Lydia Bremer-McCollum, Nicholas Wagner, and Amir Zeldes. 2024. Lacuna language learning: Leveraging RNNs for ranked text

completion in digitized Coptic manuscripts. In *Proceedings of the 1st Workshop on Machine Learning for Ancient Languages (ML4AL 2024)*, pages 61–70, Hybrid in Bangkok, Thailand and online. ACL.

Michael Paul, Andrew Finch, and Eiichrio Sumita. 2013. How to choose the best pivot language for automatic translation of low-resource languages. *ACM Transactions on Asian Language Information Processing (TALIP)*, 12(4):1–17.

Amy Pu, Hyung Won Chung, Ankur P Parikh, Sebastian Gehrmann, and Thibault Sellam. 2021. Learning compact metrics for mt. In *Proceedings of EMNLP*.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. ACL.

Muhammed Saeed, Asim Mohamed, Mukhtar Mohamed, Shady Shehata, and Muhammad Abdul-Mageed. 2024. From nile sands to digital hands: Machine translation of coptic texts. In *Proceedings of The Second Arabic Natural Language Processing Conference*, pages 298–308.

Aleksi Sahala and Eliese-Sophia Lincke. 2025. Neural models for lemmatization and POS-tagging of earlier and late Egyptian (supporting hieroglyphic input) and demotic. In *Proceedings of the Second Workshop on Ancient Language Processing*, pages 99–104, The Albuquerque Convention Center, Laguna. ACL.

Caroline T Schroeder and Amir Zeldes. 2016. Raiders of the lost corpus. *Digital Humanities Quarterly*, 10(2).

Thibault Sellam, Dipanjan Das, and Ankur P Parikh. 2020. Bleurt: Learning robust metrics for text generation. In *Proceedings of ACL*.

Jörg Tiedemann, Mikko Aulamo, Daria Bakshandaeva, Michele Boggia, Stig-Arne Grönroos, Tommi Nieminen, Alessandro Raganato, Yves Scherrer, Raul Vazquez, and Sami Virpioja. 2023. Democratizing neural machine translation with OPUS-MT. *Language Resources and Evaluation*, (58):713–755.

Christina Tzogka, Fotini Koidaki, Stavros Doropoulos, Ioannis Papastergiou, Efthymios Agrafiotis, Katerina Tiktopoulou, and Stavros Vologiannidis. 2021. Ocr workflow: Facing printed texts of ancient, medieval and modern greek literature.

Amir Zeldes and Caroline T Schroeder. 2016. An nlp pipeline for coptic. In *Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 146–155.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.