SampoNLP：用于对子词分词器进行形态分析的自指工具包

Iaroslav Chelombitko
DataSpike，aglabx 帕福斯尼亚波利斯大学 帕福斯，塞浦路斯
i.chelombitko@nup.ac.cy

Ekaterina Chelombitko
DataSpike 迪拜，阿联酋
ekaterina@dataspike.io

Aleksey Komissarov
aglabx 帕福斯，塞浦路斯
ad3002@gmail.com

arXiv:2601.04469v1 [cs.CL] 8 Jan 2026

摘要

子词分词的质量对大型语言模型至关重要，但由于缺乏干净的语素词典，评估形态丰富的乌拉尔语系语言的分词器仍面临阻碍。

我们推出 SampoNLP，这是一款用于构建形态词库的无语料库工具包，采用受最小描述长度启发的自指原子性评分，通过内部结构线索筛选复合形式，适用于低资源环境。

我们使用 SampoNLP 为芬兰语、匈牙利语和爱沙尼亚语生成的高纯度词典，对不同词汇表大小（8k–256k）下的 BPE 分词器进行了系统评估。我们提出了一种统一指标——综合性能得分（IPS），用于权衡语素覆盖率与过度切分。通过分析 IPS 曲线，我们确定了收益递减的“肘点”，并首次为这些语言的最佳词汇表大小（k）提供了基于实证的建议。本研究不仅提供了实用指导，还定量展示了标准 BPE 在高度黏着语中的局限性。SampoNLP库及所有生成的资源均已公开提供[1]。

1 引言

字节对编码（BPE）等子词分词算法的表现（Sennrich 等人，2016）是现代自然语言处理（NLP）的基石。尽管这些算法对许多语言都非常有效，但其纯统计特性给形态丰富、黏着性强的语言带来了重大挑战（Bostrom 和 Durrett，2020；Rust 等人，2021）。在乌拉尔语系中，这类语言以复杂的形态结构和多样的语言现象著称（Hämäläinen，2019），词语往往是由多个语素串接而成的长词（例如，芬兰语 talo-i-ssa-ni-ko-kaan——“我的房子也不在里面吗？”）。对于这类语言，分词质量不仅是工程细节，更是决定模型能否掌握语法结构并有效泛化的关键因素（Hämäläinen 等人，2021；Gerz 等人，2018年）。这引出了一个迫切但尚未得到充分探讨的实际问题，也是乌拉尔语系自然语言处理领域公认的挑战：要实现稳健的形态表征，最佳的分词器词汇表大小（k）是多少？近期研究展示了专为这些语言设计的分词器所带来的优势，凸显了这一问题的重要性（Chelombitko 和 Komissarov，2024）。

回答这一问题揭示了一个更根本的问题：用于评估的高纯度形态学资源十分匮乏。尽管拼写检查词典中提供了词汇数据，但将词干和词缀直接组合会产生噪声较大的候选列表。人工整理无法扩展，而 Morfessor 等成熟的基于语料库的方法（Creutz 和 Lagus，2007）并不适用于众多低资源乌拉尔语系语言（Arkhangelskiy，2019）。

为应对这一挑战，我们提出了 SampoNLP，这是一款基于无语料库、自指式流程的工具包，用于完善形态学词典。所提出的方法“受最小描述长度启发的自指式原子性评分”，其理论动机源自最小描述长度原理（Rissanen，1978年），但将其调整为仅类型设置。核心算法通过分析数据集本身的内部结构模式，迭代估算每个候选项的原子性，并区分简单形式与复合形式。这种轻量且可复现的方法为生成更完善的形态学资源提供了一种实用途径；对于数据稀缺环境而言，这类资源是公认的需求，而传统的基于语料库的方法在这些环境中并不可行（Hämäläinen，2019）。

[1] https://github.com/AragonerUA/SampoNLP

在建立了稳健的资源创建方法论后，我们利用生成的词典解决本文的核心问题：BPE 分词固有的词汇量—形态学权衡（Bostrom 和 Durrett，2020）。我们对芬兰语、匈牙利语和爱沙尼亚语的 BPE 分词器进行了系统评估，词汇表规模涵盖 8k 至 256k。开发超越下游性能的新型评估框架正日益成为一个重要研究领域（Chelombitko 等人，2024）。与此一致，为了精确把握上述权衡，我们提出综合性能分数（IPS），这一单一指标能够平衡词汇语素覆盖率（LMC）与过度切分率（OSR）。借助该指标，我们可以对性能曲线建模并确定最佳词汇量范围，从而为我们的核心研究问题提供有据可循的答案。

因此，我们的贡献有两方面，且同等重要：

1. **无需语料库的形态学方法：** 我们推出了一套完全自动化且可复现的流程，无需依赖语料频率或外部资源即可优化形态学词典，并以开源工具包 *SampoNLP* 的形式发布。

2. **定量评估：** 我们对芬兰语、爱沙尼亚语和匈牙利语的 BPE 分词器进行了系统分析，通过新定义的覆盖率和过度切分指标，考察词汇表大小如何影响形态学粒度。

2 相关工作

对形态丰富语言的子词分词进行评估和优化，涉及多个研究领域：子词分词算法、无监督形态学分析、基于规则的分析器，以及面向乌拉尔语系语言的特定语言自然语言处理。

2.1 子词分词与形态学

字节对编码（BPE）（Sennrich 等人，2016）已成为现代自然语言处理中子词分词的事实标准。除了这一方法，人们还提出了单元语言模型（Kudo，2018年）等方法，但这些方法纯粹的统计特性给形态复杂语言带来了众所周知的挑战。（Bostrom 和 Durrett，2020）的研究表明，BPE 分词器往往无法与语言学语素边界对齐。有趣的是，在自然语言处理之外的领域中，人们也探讨了识别有意义的子序列单元所面临的类似挑战，例如对灵长类基因组等生物序列进行分词（Popova 等人，2025年）。

最佳词汇表大小的问题，通常依赖启发式方法来确定，或通过下游任务性能间接评估（Mielke 等人，2021年）。我们的工作通过提出一种用于内在且以形态学为基础的评估方法，直接弥补了这一空白，从而为乌拉尔语系语言提供数据驱动的建议。

2.2 无监督形态学分析

无监督发现形态结构有着悠久的历史。一大类方法依靠语料库中的统计线索来识别边界。分支熵和接续符多样性等经典方法（Chen 等人，2004）通过分析后续字符的可预测性来推测语素边界。另一类重要方法则基于最小描述长度（MDL）原理。Morfessor（Creutz 和 Lagus，2007）及其变体是典型的概率方法，通过寻找能够最佳压缩文本语料库的词典来实现分析。尽管这些方法效果良好，但从根本上说，它们依赖语料库，需要词元频率信息，而这类信息在低资源环境中可能无法获得。

我们的方法虽然受到最小描述长度的启发，但采用的是无语料库、仅词型模式。它代表了一种不同的范式：对候选列表进行自指式筛选。通过完全依据候选集合的内部结构开展工作，我们提供了一种轻量级方法，适用于资源匮乏的场景；这一直是乌拉尔语系自然语言处理面临的挑战（Arkhangelskiy，2019）。

2.3 乌拉尔语系语言的基于规则的分析器与分词

对于乌拉尔语系语言而言，基于有限状态转换器（有限状态转换器）构建的规则型形态分析器，例如 Omorfi（Pirinen，2015年）以及 GiellaLT[2] 基础设施（Jauhiainen 等人，2020），都是极其宝贵的资源。尽管它们生成的结果在语言学上十分全面，却并未针对用作精简的参考语素词典而直接优化。我们的 IMDP 流程提供了一种不同的

[2]https://giellalt.github.io/

方法：采用数据驱动方法，从仅词型候选列表中提炼出此类词典。此类候选列表可从 Hunspell 等基于词典的资源中提取，且无需语料库中的词元频率。

近年来，如何为这一语系实现有效分词已成为备受关注的挑战。更广泛的研究发现表明，对于形态丰富的语言，采用特定语言的建模方式至关重要；芬兰语研究显示，与多语模型相比，FinBERT 等单语模型具有明显优势（Virta- nen 等人，2019）。基于这一原则，Chelombitko和Komissarov（2024）最近的一项研究专门探讨了乌拉尔语系语言在大型多语模型中的严重代表性不足问题。他们证明，训练专用的大型词汇表单语分词器可以显著提高压缩效率。然而，尽管他们的研究指出了专用资源的必要性，却将如何确定最佳词汇表大小这一问题留待未来研究。

与此同时，对更优评估指标的需求已成为备受关注的研究课题。例如，Qtok 框架（Chelombitko 等人，2024）提出了一种全面评估多语分词器质量的方法；其他研究也主张超越下游任务性能，转向更具内在性且具有语言学依据的衡量方式（Beinborn和Pinter，2023）。我们的综合性能得分（IPS）直接回应了学界对更多基于形态学的指标的呼吁。

当前工作建立在这些基础之上。我们使用与（Chelombitko和Komissarov，2024）中相似的高质量数据源来训练待评估的分词器。此外，通过提出一套具体的方法论，我们响应了改进评估的呼吁，并找到了前一项研究所暗示的最优词汇表规模，从而为这一研究方向提供了合乎逻辑的下一步。

3 方法论。IMDP 流程

为了从含噪的原始候选形式列表中创建高纯度语素词典，我们提出了迭代形态分解流程（IMDP）。我们的方法旨在实现全自动化，并在无语料库、仅词型模式下运行，只需要候选列表作为输入。该流程的核心是一种我们称为“受最小描述长度启发的自指原子性评分”的方法，它通过迭代评估每个候选项相对于整个集合的“基础性”。整个过程如图1所示。

该流程包括三个主要阶段：(1) 预过滤和初始评分，(2) 迭代式评分细化，以及 (3) 通过自动阈值处理进行最终过滤。

3.1 阶段 1：候选项预过滤和初始评分

这一初始阶段旨在大幅减少非语言噪声，并为每个合理的候选项建立基线评分。

3.1.1 硬性预过滤

首先，我们对原始输入列表 C_{raw} 应用一系列确定性过滤器。如果词元 $t \in C_{raw}$ 满足以下任一条件，则将其丢弃：

1. 包含来自非目标文字的符号（例如，拉丁字母列表中的西里尔文）。我们为每种语言定义一个有效字符集 Σ（例如，匈牙利语使用 [a-záéíóö˝oúü˝u]）。

2. 包含任何非字母字符（例如数字、标点符号、网址），但用于标记词缀的开头/结尾连字符除外。

3. 属于专有名词或首字母缩略词（启发式判断：以大写字母开头，或由多个大写字母组成）。

4. 过长（$|t| > 30$）或过短（$|t| < min_length$），除非 t 是一个出现在特定语言有效单字符语素白名单 W 中的单字符。

3.1.2 类型支持过滤

为了过滤掉排版错误和其他单词元噪声，我们对剩余候选项集合 C'应用“类型支持”准则。只有当候选项 $t \in C'$作为子字符串出现在 C'中至少 m个其他唯一候选项中时，才予以保留。这确保我们只考虑在数据集本身中反复出现的结构模式。$support(t) = |\{c \in C' | t是\ c\}|$ 的子字符串。若满足 $support(t) \geq m$，我们就保留 t（我们使用 $m = 3$）。最终得到的集合即为候选池 C。

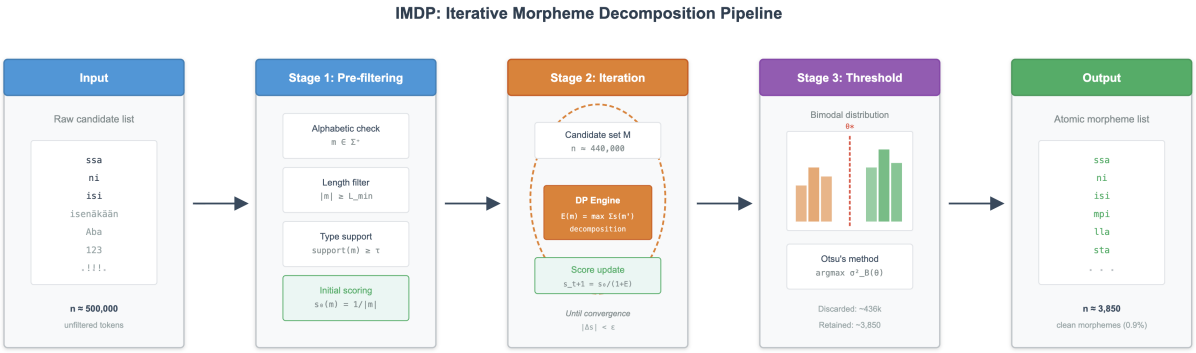

图1：迭代形态分解流程（IMDP）概览。

3.1.3 初始原子性评分

每个保留下来的候选项 $t \in C$ 都会被赋予一个初始原子性分数 $S_0(t)$。该分数基于最小描述长度启发的原则：在其他条件相同的情况下，较短的形式更可能是基本语素单位。分数定义为词元长度的倒数：$S_0(t) = \frac{1}{|t|}$，其中 $|t|$ 是 t中的字符数。

3.2 阶段二：迭代分数优化

这是我们方法的核心。我们会迭代优化原子性分数，直到分数收敛。在每次迭代 $k+1$ 中，每个词元 $t \in C$ 的分数都会根据集合中其他词元对它的“可解释性”重新计算。

3.2.1 最优分解与最佳解释力（BEP）

对于每个词元 t，我们将其最优分解为一串更小的词元（m_1，m_2，..... m_n），其中每个 $m_i \in C$。最优分解是指其组成部分的分数之和（取自上一次迭代，S_k）最大的分解。我们使用动态规划算法求出这个最大和，并将其称为最佳解释力，$BEP_k(t)$。

$$\mathrm{BEP}_k(t) = \max_{\substack{t=m_1\cdots m_n \\ n\geq 2}} \sum_{i=1}^{n} S_k(m_i).$$

分解的搜索空间受以下两条规则限制：

1. **多组件：** 算法会考虑分成任意数量部分的切分，而不局限于两部分。

2. **退化预防：** 只有当长度为 1 的片段位于白名单 W 中时，才会考虑它们。

3.2.2 分数更新规则

新得分 $S_{k+1}(t)$ 的计算方式是将词元自身的得分与其可解释性进行比较。仅当表明该词元为复合词元的“证据”（$BEP_k(t)$）强于表明其为原子词元的证据（$S_k(t)$）时，才会对该词元进行惩罚。

$$S_{k+1}(t) = \begin{cases} S_k(t), & \text{if } \mathrm{BEP}_k(t) \leq S_k(t), \\ \dfrac{S_0(t)}{1+\mathrm{BEP}_k(t)}, & \text{if } \mathrm{BEP}_k(t) > S_k(t). \end{cases}$$

此更新规则形成了一种竞争动态：原子语素保留较高分数，而复合词则在迭代过程中不断受到惩罚，分数逐渐趋近于零。

3.2.3 收敛

迭代过程会持续进行，直到系统达到稳定状态。我们将收敛定义为：连续两次迭代之间，任意词元分数的最大绝对变化低于一个较小阈值的时刻

$$\max_{t\in\mathcal{C}} \left| S_{k+1}(t) - S_k(t) \right| < \varepsilon$$

我们使用 $\varepsilon = 1e-7$，并设置 $max\ iterations = 100$的保护上限。

3.3 阶段3：通过自动阈值处理进行最终筛选

分数收敛后，最终的分数分布通常会显示出这样的特征：复合候选项高度集中在极低分数区间，而原子候选项则保持较高分数。为了自动且可复现地确定这两组之间的分离阈值，我们采用大津法（Otsu，1979）。该算法最初用于图像处理，以分离前景和背景，通过寻找最优阈值 τ

，即通过最大化两个所得类别之间的类间方差来划分数据分布（在我们的案例中为“原子”与“复合”）。这种数据驱动的方法避免了手动参数调优，并能适应每个数据集特有的分数分布。

所有最终得分为 $S_{final}(t) >= \tau$的词元 t 都被归类为原子，并构成我们最终的高纯度语素词典。

Lang	初始候选项	原子形态素	约简%	约简因子
Fin	499,647	3,850	99.23%	129.8 倍
Est	281,256	5,705	97.97%	49.3 倍
Hung	103,317	3,189	96.91%	32.4x

表 1：IMDP 流程在清理和缩减语素候选列表方面的效率。

4 实验设置

为评估词汇表大小对形态覆盖率的影响，我们对三种乌拉尔语系语言进行了系统分析：芬兰语、匈牙利语和爱沙尼亚语。我们的实验设置包括三个主要阶段：创建参考语素、训练分词器以及定义评估指标。

4.1 数据

我们的方法论要求为每种语言准备两类数据：用于清理的原始语素候选项列表，以及用于分词器训练的大型文本语料库。

1.**语素候选列表：**最初的“脏”候选列表是根据基于 Hunspell 框架的权威开源拼写检查词典构建的[3]。对于匈牙利语和爱沙尼亚语，我们使用了 LibreOffice项目整理的综合词典[4]。对于需要特别处理复合词的芬兰语，我们使用了 hunspell-fi项目提供的专用词典[5]。对于每种语言，我们将所有不重复的词干（来自 .dic 文件）和词缀（来自 .aff 文件）合并，生成一份全面但结构嘈杂的候选列表，作为 IMDP 流程的输入。这种方法利用广泛可用的词典资源，为形态分析提供了切实可行的起点。

2. **文本语料库：** 在训练 BPE 分词器时，我们使用了从 Wikipedia 快照中提取的大型预处理语料库[6]。我们对数据源和预处理方法论的选择与此前创建专用乌拉尔语系语言分词器的工作保持一致（Chelombitko和 Komissarov，2024），从而为我们的分析确保了可比的基础。必须强调的是，这些语料库仅用于训练 BPE 分词器，并未用于语素列表优化流程的任何阶段，因此保留了 IMDP 方法不依赖语料库的特性。

4.2 参考语素词典创建

对于三种语言中的每一种，我们都按照第 3 节所述，针对相应的原始候选列表应用了迭代形态分解流程（IMDP）。该流程配置了以下参数：最小语素长度 $min_length = 1$、最小类型支持度 $m = 3$和收敛阈值$\varepsilon = 1e-7$。该过程一直运行到收敛。最终过滤使用自动确定的大津阈值（Otsu，1979）完成。该流程生成了三个高纯度参考语素词典（G_{fin}、G_{hun}、G_{est}），其统计数据汇总于表1中。

4.3 分词器训练

我们使用 tokenizers 库[7] 和 SentencePiece（Kudo 和 Richardson，2018年）进行比较，并从头开始为每种语言训练了一系列字节对编码（BPE）分词器。这些分词器在各自的维基百科语料库上进行训练。为分析词汇表大小的影响，我们针对一系列词汇表大小 k 分别训练了模型，范围从 8,000 到 256,000（$k \in${8k, 16k, 32k, 40k, 50k, 64k, 80k, 100k, 128k, 150k, 180k, 200k, 220k, 240k, 256k}）。所有分词器的合并操作均设置为 $min_frequency$，合并操作数量为 2。

4.4 评估指标

为了对分词器质量进行细致而严谨的评估，我们必须考虑形态覆盖率与

[3]https://hunspell.github.io/[4] https://github.com/LibreOffice/dictionaries[5] https://github.com/fginter/hunspell-fi

[6]https://dumps.wikimedia.org[7] https://github.com/huggingface/tokenizers

过度切分之间的基本权衡。能够完美表示所有语素（覆盖率高）的分词器，如果也会过度拆分常见词，就并非最优。为了用一个统一的分数体现这种平衡，我们引入综合性能得分（IPS）。

IPS 以几何方式对这种权衡进行建模。我们考虑一个二维空间，其中理想的分词器位于点 (Coverage=1,OverSplit=0)。任何实际分词器的 IPS，都是它与该理想点之间的归一化欧几里得距离，并缩放到 $[0, 1]$ 范围，其中 1 表示完美。

首先，我们定义两个核心组成部分：

1. **词汇语素覆盖率（LMC）：** 我们的参考词典中原子语素所占的比例：这些语素 G 被分词器的词汇表 V_k完美地表示为单个词元。这衡量了分词器对基本形态单位的词汇“认知”。

$$\text{LMC} = \frac{\left|\{ m \in G \mid m \in V_k \}\right|}{|G|}.$$

2. **过度切分率（OSR）：** G 中未能被分词器表示为单个词元、因而总是被拆分成多个片段的语素所占的比例。

$$\text{OSR} = \frac{\left|\left\{ m \in M \,\middle|\, \begin{array}{l} m \text{ occurs in } \geq 1 \text{ word} \\ m \text{ never as a single token} \end{array} \right\}\right|}{|\{ m \in M \mid m \text{ in } \geq 1 \text{ word} \}|}.$$

综合性能得分（IPS）的计算方式如下：

$$IPS = 1 - \left(\frac{\sqrt{(1-LMC)^2 + OSR^2}}{\sqrt{2}}\right)$$

这一单一指标便于清晰、直接地比较不同词汇表大小下的分词器。IPS 越高，表示在表示语素与避免过度碎片化之间的平衡越好。我们最终通过确定 IPS 随词汇表大小变化曲线上的“肘点”，来分析最佳词汇表大小。

5 结果与分析

我们的实验得出了清晰且显著的模式，揭示了分词器词汇表大小与形态学性能之间的关系。为了刻画覆盖率与过度切分之间的基本权衡，我们分析了每种语言的综合性能得分（IPS）。爱沙尼亚语（图 5）、芬兰语（图 6）和匈牙利语（图 4）的 IPS 曲线清晰地展示了每种语言的性能概况。组件指标（LMC 和 OSR）的补充详情见图 2和3。

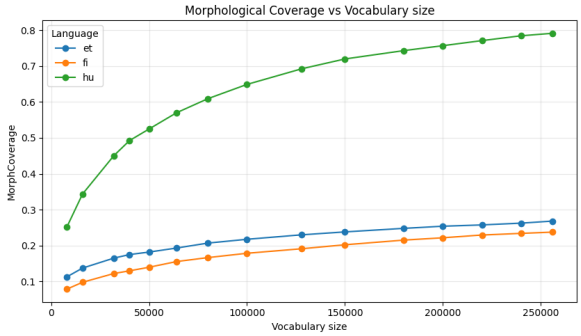

图 2：不同词汇表规模（k）下的词汇语素覆盖率（LMC）。LMC 表示参考语素中可作为完整独立词元出现在分词器词汇表中的比例。

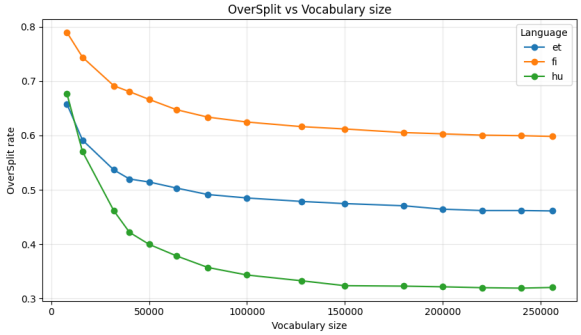

图 3：过度切分率（OSR）随词汇表规模（k）的变化。OSR 表示参考语素中出现在词语里、但在任何分词结果中都不会作为单个词元出现的比例。

5.1 总体观察：清晰的权衡特征

这三种语言的 IPS 曲线均呈现出典型的对数增长模式，体现了边际收益递减规律。词汇表大小较小时，得分增长迅速，表明最初增加词汇表内容能高效地捕捉形态结构。然而，改进速度逐渐放缓，说明不断扩大词汇表只能带来微小收益，却会显著增加模型大小。这证实了每种语言都存在一个“最佳区间”或最佳范围。

5.2 跨语言分析：三个不同的性能等级

结果揭示了三个截然不同的性能等级，凸显了标准字节对编码对这些语言形态学的建模能力各不相同。

1. **匈牙利语（hu）：** 如图 4所示，匈牙利语的表现远胜其他语言。其 IPS 曲线从 0.29 起步，随后急剧上升，最高达到 0.73。如此高的得分表明，字节对编码在学习匈牙利语形态学的统计规律方面颇为有效。

2. **爱沙尼亚语（et）：** 爱沙尼亚语处于中等水平，其 IPS 曲线如图 5所示。得分从 0.22 起步，最高达到 0.39。虽然优于芬兰语，但这一得分表明，即使词汇表很大，分词器的表现也未达到“理想”水平的 40%。

3. **芬兰语（fi）：** 图 6 展示了芬兰语最具挑战性的特征。IPS 最高仅为 0.31，结果以量化方式表明，标准字节对编码根本不适合捕捉芬兰语形态学的复杂性。

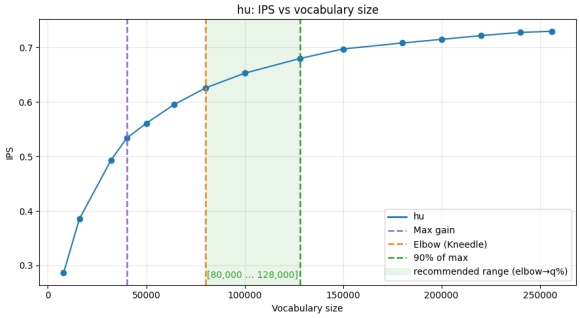

图4：匈牙利语的 IPS 与词汇表大小（k）的关系。匈牙利语的 IPS 提升最为稳定，这反映了其黏着结构相对透明，形态音韵交替也较少。肘点位于 80k，90% 质量阈值位于 128k，因此推荐范围为 8万–12.8万。

5.3 确定最佳词汇量范围（k*）

为了确定实用且有效的词汇表大小，我们为 k* 定义了一个推荐范围。该范围的下限是“肘点”(k_elbow)，由 Kneedle算法（Satopää等人，2011年）确定，标志着边际收益递减的起点。上限是 90% 质量点（k_q90），即达到观测到的最高 IPS 的 90% 时对应的点。如图4、6、5所示，并汇总于表 2，这项分析得出了以下建议：

1. **匈牙利语（hu）：** 匈牙利语的 IPS 曲线（图4）显示，最佳范围明确位于 k=80,000 到 k=128,000 之间。肘点位于 80k，达到最高性能的 90% 时对应的词汇表大小为 128k。从图中可以看出，超出此范围继续扩大词汇表，带来的性能提升微乎其微。

2. **爱沙尼亚语（et）：** 对于爱沙尼亚语（图5），推荐范围同样为 k=80,000 到 k=128,000。与匈牙利语类似，肘点位于 80k，90% 质量阈值位于 128k，因此这一区间是在性能与规模之间取得最佳平衡的范围。

3. **芬兰语（fi）：** 芬兰语的分析（图6）表明需要更大的词汇表。拐点位于 k=80,000，但若要达到最高性能（尽管最高性能本身不高）的 90%，则需要 k=150,000 的词汇表。这表明，对于芬兰语，最佳范围为 k=80,000 至 k=150,000，反映了该语言较高的形态复杂性。

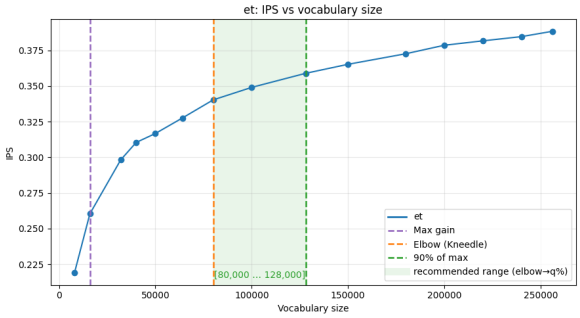

图5：爱沙尼亚语的 IPS 与词汇表大小（k）。尽管总体上的边际收益递减模式与匈牙利语相似，但 IPS 平台期较低，表明爱沙尼亚语大量的形态音系交替降低了可学习性，并掩盖了正字法中的语素边界。推荐范围仍为 8万–12.8万。

这些发现为词汇表规模设定这一关键决策提供了定量基础，使其从基于启发式的选择转变为有原则的优化问题。完整的数值

Lang	最大增益点 (k_gain)	肘点 (k_elbow)	90% 质量点 (k_q90)	推荐 k* 范围
Hung	40,000	80,000	128,000	80k–128k
Est	16,000	80,000	128,000	80k – 128k
Fin	64,000	80,000	150,000	80k – 150k

表2：用于确定最佳词汇量范围的 IPS 曲线关键点。

结果列于附录A（表3），供参考。

6 结论

在本研究中，我们应对了双重挑战：在无语料库的环境下创建高纯度形态学资源，并利用这些资源评估适用于乌拉尔语系语言的子词分词器。我们推出了 SampoNLP，这是一个工具包，采用基于“受最小描述长度启发的自指原子性评分”的新颖流程，成功地将嘈杂的候选列表提炼为干净的语素词典。

利用这些词典，我们对 BPE 分词器进行了系统评估，并得出了两项关键发现。首先，我们基于实证提出了最佳词汇表大小建议：匈牙利语和爱沙尼亚语的范围为 80k–128k，芬兰语为 80k–150k；这些范围在性能与模型大小之间实现了最有效的权衡。其次，我们的结果定量地证明了标准 BPE 对芬兰语等高度黏着语的严重局限性：其性能停滞在低得惊人的水平。

本研究证实，尽管优化词汇表大小是至关重要的一步，但它并非万能良方。我们将 SampoNLP 库和生成的语素列表公开给社区，以促进可复现研究，并鼓励开发针对乌拉尔语系、具备更多形态学意识的分词方法。

讨论

我们的结果揭示了两个关键发现。首先，字节对编码的效用因语言而异，差异显著：匈牙利语的 IPS 较高（最高为 ~0.73），而芬兰语（~0.31）和爱沙尼亚语（~0.39）的低分则从量化角度证明了该算法在这些高度黏着语上的根本局限。其次，对于所有语言，都存在一个可通过经验识别的词汇表大小“最佳区间”，超过该区间后，性能提升会逐渐减弱。此处，“最优性”指形态学上的充分性——即分词器能够捕捉语言的能产结构，同时将冗余降至最低的临界点。

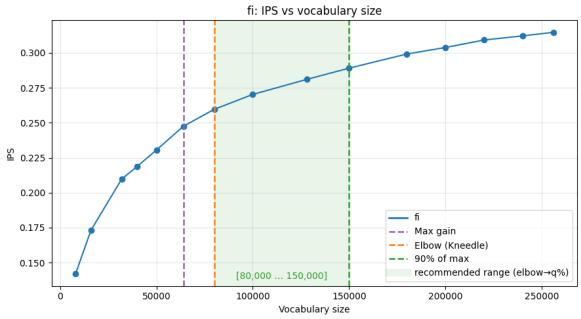

图6：**芬兰语**的 IPS 与词汇表大小（k）的关系。芬兰语的 IPS 平台值最低，这与其丰富的辅音交替系统和词干变化相符；这些特征使字节对编码的正字法切分稳定性较差。拐点位于 80k，而达到最大 IPS 的 90% 需要 150k，这表明建议范围为 80k–150k。

这一概念是设计本身所固有的，它提供的是语言层面的评判标准，而非针对特定任务的优化标准。

我们承认本方法存在局限。IPS 指标忽略了切分在定性方面的差异——这是为实现可扩展性而做出的必要取舍。我们使用干净、标准化的语料库，也隔离了词汇表大小这一变量，但未能反映真实数据中的噪声。这些方面为未来研究指明了明确方向。

虽然我们的方法生成了一组经过优化的重复出现的词汇下单位，但我们并不声称这些单位在语言学上完全符合语素划分。IMDP 分词本质上是正字法层面的自指过程，提供的是一种实用的近似，而非以音系学为基础的形态分析。

总之，我们的研究结果表明，尽管优化 k* 是关键步骤，但对于芬兰语等语言而言，这可能还不够。字节对编码的性能上限较低，这凸显了形态学感知分词方法的必要性。我们相信，SampoNLP 工具包和生成的词典将为社群提供一个可复现的基准，以开发和测试此类新策略。

参考文献

Timofey Arkhangelskiy.2019. 乌拉尔语系少数民族语言的社交媒体语料库。载于第五届乌拉尔语系语言计算语言学国际研讨会论文集，第 125–140 页，爱沙尼亚塔尔图。计算语言学协会。

Lisa Beinborn 和 Yuval Pinter。2023。分析子词分词的认知合理性。载于*2023*年自然语言处理实证方法会议论文集，第 4478–4486 页，新加坡。计算语言学协会。

Kaj Bostrom 和 Greg Durrett。2020。字节对编码不适用于语言模型预训练。载于计算语言学协会论文集：*EMNLP 2020*，第 4617–4624 页，线上举行。计算语言学协会。

Iaroslav Chelombitko 和 Aleksey Komissarov。2024。用于大型语言模型中乌拉尔语系语言表征的专用单语BPE分词器。载于第九届乌拉尔语系计算语言学国际研讨会论文集，第 89–95 页，芬兰赫尔辛基。计算语言学协会。

Iaroslav Chelombitko、Egor Safronov 和 Aleksey Komissarov。2024年。Qtok：用于评估大型语言模型中多语分词器质量的综合框架。预印本，arXiv:2410.12989。

Kang Chen、Xiaotie Deng 和 Weimin Zheng。2004年。用于中文词语抽取的接续多样性标准。《计算语言学》，30:75–93。

Mathias Creutz 和 Krista Lagus。2007年。用于语素切分和形态学习的无监督模型。*ACM* 语音与语言处理汇刊，4(1)。

Daniela Gerz、Ivan Vuli´c、Edoardo Ponti、Jason Naradowsky、Roi Reichart 和 Anna Korhonen。2018年。形态丰富语言的语言建模：用于词级预测的字符感知建模。《计算语言学协会汇刊》，6:451–465。

Mika Hämäläinen、Niko Partanen、Jack Rueter 和 Khalid Alnajjar。2021年。面向多种语言的神经形态学数据集与模型：从大型语言到濒危语言。载于《第二十三届北欧计算语言学会议论文集（*NoDaL- iDa*）》，第 166–177 页，冰岛雷克雅未克（线上）。瑞典林雪平大学电子出版社。

Mika Hämäläinen。2019。Uralicnlp：面向乌拉尔语系语言的自然语言处理库。《开放源代码软件期刊》，4(37)。

T. Jauhiainen、Krister Linden、Niko Partanen 等。2020年。乌拉尔语系语言识别（uli）2020年共享任务：用于乌拉尔语的 Wanca 2017网络语料库。*LREC 2020 VarDial*研讨会论文集。

Taku Kudo。2018年。子词正则化：利用多个子词候选项改进神经网络翻译模型。载于计算语言学协会第*56*届年会论文集（第*1*卷：长篇论文），第66–75页，澳大利亚墨尔本。计算语言学协会。

Taku Kudo 和 John Richardson。2018年。SentencePiece：一种用于神经文本处理的简单、语言无关的子词分词器和反分词器。载于《*2018*年自然语言处理实证方法会议：系统演示论文集》，第 66–71 页，比利时布鲁塞尔。计算语言学协会。

Sabrina J. Mielke、Zaid Alyafeai、Elizabeth Salesky、Colin Raffel、Manan Dey、Matthias Gallé、Arun Raja、Chenglei Si、Wilson Y. Lee、Benoît Sagot 和 Samson Tan。2021年。词语与字符之间：开放词汇建模与自然语言处理分词简史。预印本，arXiv:2112.10508。

Nobuyuki Otsu。1979。基于灰度直方图的阈值选择方法。*IEEE*《系统、人类与控制论汇刊》，9(1):62–66。

Tommi A Pirinen。2015年。Omorfi——面向芬兰语的免费开源形态词汇数据库。载于第*20*届北欧计算语言学会议论文集（*NODALIDA 2015*），第313–315页，立陶宛维尔纽斯。林雪平大学电子出版社，瑞典。

Marina Popova、Iaroslav Chelombitko 和 Aleksey Komissarov。2025年。重复如何驱动词汇表：对端粒到端粒灵长类基因组进行字节对编码分析。预印本，arXiv:2505.08918。

Jorma Rissanen。1978年。通过最短数据描述建模。《自动化》，14(5):465–471。

Phillip Rust、Jonas Pfeiffer、Ivan Vuli´c、Sebastian Ruder 和 Iryna Gurevych。2021年。你的分词器有多好？多语言模型的单语性能。载于第*59*届计算语言学协会年会暨第*11*届自然语言处理国际联合会议论文集（第*1*卷：长篇论文），第3118–3135页，线上举行。计算语言学协会。

Ville Satopää、Joshua Albrecht、David Irwin 和 Barath Raghavan。2011年。大海捞针：检测系统行为中的膝点。载于第*31*届分布式计算系统国际会议研讨会论文集，第166–171页。IEEE。

Rico Sennrich、Barry Haddow 和 Alexandra Birch。2016 年。使用子词单元进行罕见词的神经机器翻译。载于：第 *54* 届计算语言学协会年会论文集（第*1*卷：长篇论文），第 1715–1725 页，德国柏林。计算语言学协会。

AnttiVirtanen、Jenna Kanerva、Rami Ilo、Jouni Luoma、Juhani Luotolahti、Tapio Salakoski、Filip Ginter 和 Sampo Pyysalo。2019 年。多语言还不够：用于芬兰语的 BERT。预印本，arXiv：1912.07076。

A 附录

语言	词汇表规模（千）	总计语素	语素覆盖率 %	过度切分率 %
爱沙尼亚语	8,000	5,705	11.27%	65.79%
爱沙尼亚语	16,000	5,705	13.71%	59.13%
爱沙尼亚语	32,000	5,705	16.49%	53.65%
爱沙尼亚语	40,000	5,705	17.48%	51.98%
爱沙尼亚语	50,000	5,705	18.18%	51.43%
爱沙尼亚语	64,000	5,705	19.30%	50.32%
爱沙尼亚语	80,000	5,705	20.68%	49.13%
爱沙尼亚语	100,000	5,705	21.74%	48.49%
爱沙尼亚语	128,000	5,705	22.99%	47.86%
爱沙尼亚语	150,000	5,705	23.79%	47.46%
爱沙尼亚语	180,000	5,705	24.78%	47.06%
爱沙尼亚语	200,000	5,705	25.38%	46.43%
爱沙尼亚语	220,000	5,705	25.74%	46.19%
爱沙尼亚语	240,000	5,705	26.23%	46.19%
爱沙尼亚语	256,000	5,705	26.81%	46.11%
芬兰语	8,000	3,850	7.85%	78.96%
芬兰语	16,000	3,850	9.76%	74.37%
芬兰语	32,000	3,850	12.20%	69.13%
芬兰语	40,000	3,850	12.95%	68.04%
芬兰语	50,000	3,850	13.95%	66.62%
芬兰语	64,000	3,850	15.53%	64.73%
芬兰语	80,000	3,850	16.63%	63.36%
芬兰语	100,000	3,850	17.84%	62.46%
芬兰语	128,000	3,850	19.11%	61.61%
芬兰语	150,000	3,850	20.21%	61.18%
芬兰语	180,000	3,850	21.51%	60.52%
芬兰语	200,000	3,850	22.16%	60.28%
芬兰语	220,000	3,850	22.93%	60.05%
芬兰语	240,000	3,850	23.38%	59.95%
芬兰语	256,000	3,850	23.73%	59.81%
匈牙利语	8,000	3,189	25.15%	67.72%
匈牙利语	16,000	3,189	34.34%	57.01%
匈牙利语	32,000	3,189	45.03%	46.14%
匈牙利语	40,000	3,189	49.23%	42.17%
匈牙利语	50,000	3,189	52.46%	39.97%
匈牙利语	64,000	3,189	56.98%	37.84%
匈牙利语	80,000	3,189	60.90%	35.72%
匈牙利语	100,000	3,189	64.88%	34.33%
匈牙利语	128,000	3,189	69.24%	33.27%
匈牙利语	150,000	3,189	71.97%	32.37%
匈牙利语	180,000	3,189	74.29%	32.28%
匈牙利语	200,000	3,189	75.67%	32.16%
匈牙利语	220,000	3,189	77.08%	32.00%
匈牙利语	240,000	3,189	78.43%	31.92%
匈牙利语	256,000	3,189	79.12%	32.04%

表 3：三种乌拉尔语系语言中不同词汇表大小的 BPE 分词器的详细实验结果。语素覆盖率表示词汇表中包含的参考语素的百分比（LMC）。过度切分率表示在 W 中得到支持、但从未在任何分词结果中作为单个词元出现的参考语素的百分比。