# 将依赖关系扩展至标记 PBC：及物子句中的词序

Hiram Ring
NTU Singapore

June 10, 2025

**Abstract**

标注的 PBC（Ring 2025a ）包含了来自超过 1500 种语言的超过 1800 个句子的词性标注的平行文本数据，这些语言代表了 133 个语言家族和 111 个孤立语言。虽然这比以前可用的资源要庞大，而且词性标注达到了相当不错的准确性，允许进行预测性跨语言见解（Ring 2025b ），但该数据集最初并没有进行依赖关系标注。本文报告了一个以 CoNLLU 格式化的数据集版本，其中将依赖信息和词性标注一起传输到标注的 PBC 的所有语言。尽管关于标记和依赖关系的质量存在各种担忧，但从该数据集中衍生出的关于及物句中论元和谓词位置的信息与三个类型数据库（WALS，Grambank，Autotyp）中的专家对词序的判断相关。这突出了基于语料库的语言类型方法（如 Baylor et al. 2023; Bjerva 2024 ）的有用性，以扩展对离散语言类别的比较，并表明即使在噪声数据中也可以获得重要的见解，只要有足够的标注。依赖关系标注的语料库也通过 GitHub 对研究和合作开放。

## 1 介绍

对世界上 7000 多种语言的研究的一个重要方面是这些语言的相似性和差异。澄清这些相似性和差异一直是语言类型学的目标，各种数据库以多种方式识别和分类语言，最著名的是 WALS ( Dryer & Haspelmath 2013 ), [1] Grambank (Skirgård et al., 2023) , [2] Autotyp (Bickel et al., 2023) , [3] Glottolog ( Hammarström et al. 2024 ), [4] 和 Lexibank ( List et al. 2022 )。[5] 这些资源提供关于语言的重要信息，但常常缺乏特定类型特征和/或语言的覆盖。此外，这些数据集所做的分类判断可能不适用于所有语言，某些分类（例如词序）可能会将某种语言定型，因为特定特征可能在某些语言中被认为是边缘的或有争议的。

正如在 Ring (2025a) 中指出的，最近的发展是创建了注释数据库，通过开发强大的跨语言语料库来解决这些问题，其中最为显著的是 Universal Dependencies Treebanks 项目（UDT；Zeman et al. 2024 ）。[6] 这个数据库包含来自个别语言的语料库，并使用通用框架对词性和依赖关系进行人工注释。这样的数据集可以实现自动识别属性，从而能够按各种类型对语言进行分类或分类，反映了对计算语言类型学日益增长的兴趣（见 Baylor et al. 2023, 2024; Bjerva 2024 ），并支持不同的类型学方法（见 Levshina et al. 2023 ）。尽管像 UD

[1] https://wals.info/
[2] https://grambank.clld.org/
[3] https://github.com/autotyp/autotyp-data
[4] https://www.glottolog.org
[5] https://lexibank.clld.org/
[6] https://universaldependencies.org

这样的发展令人鼓舞，但这些数据集在覆盖范围和注释程度方面仍然面临许多与类型学数据库相同的问题。

使用计算工具开发的 taggedPBC 试图解决这些问题中的一些。如 Ring (2025a) 所指出的：

> 使用来自平行圣经语料库 (PBC; Mayer & Cysouw 2014 ) 的数据，计算词对齐工具允许在 1,597 种语言中自动标记词类（约占世界语言的 20 %），这些语言代表 133 个家族和 111 个孤立体。通过将结果与现有高资源语言的自动标注器（SpaCy, Trankit）的输出以及 UDT 中的黄金标准标注数据进行比较，此自动词性标注方法得到了验证……这些语言中的大多数（超过 1,500 种）具有超过 1,800 个平行句子或诗句，而最小的语料库则包含超过 700 个句子或诗句。

该数据集代表了最大规模的公开可用的跨语言平行语料库的标注数据库，并通过与知名类型学数据库中的词序分类的相关性建立了其有效性。正如 Ring 的研究所指出的，从这个数据集中派生出来的语料库度量（“N1 比率”）不仅能区分 WALS、Grambank 和 Autotyp 中的 SV、VS 和自由基本词序的语言，还能准确分类先前未分类语言的不及物词序。此外，在一项独立研究中，Ring (2025b) 表明从该数据集中派生的其他度量（名词和动词的平均长度）可以单独对语言的基本不及物词序进行分类。然后，这一度量被测试在不属于 taggedPBC 的手动标注语料库上，结果发现该度量甚至能够对历史变体的语言准确分类为 SV 或 VS。重要的是，Ring 的研究发现了 taggedPBC 中的趋势，这些趋势随后在金标准（手动标注）的数据集上得到验证。

虽然 Ring 的研究强调了像 taggedPBC 这样的数据集在帮助回答关于语言的跨语言问题上具有潜力，但关于标注仍有许多问题需要解决。一个主要的关注点是通过跨语言迁移所识别的词性标签的质量和覆盖范围。这体现在两个相关的问题上，即词性标签是否准确地代表了语言中的标签范围，以及相关标签是否能在英语和数据集中每种特定语言之间准确传递。

第一个相关的问题可以通过一个简单的观察来说明：并不是所有语言都包含所有语法类别。虽然所有语言似乎都包含与“名词”和“动词”相对应的词类，但是否所有语言都有“形容词”或“副词”则有很大的争议。此外，其他词类在某些特定语言中根本不存在。例如，虽然像中文这样的语言在计数表达中需要一个特殊的“量词”词类，但这一语法类别在英语中并没有真正的对应物。

这引出了第二个相关问题，即强调从英语进行跨语言标记转移的方法（如创建 taggedPBC 的方法所遵循的那样）对于某些词类比其他词类会更成功。Ring (2025a) 确实显示该方法对于名词和动词是相当成功的，但推测其他标记可能不会转移得那么好，有些标记（特别是源语言和目标语言不共享的那些）根本无法转移。就 Ring 的研究而言，趋势可能显现的原因之一（并且可以在其他数据集上进行验证）似乎是观察量相当大，这在一定程度上缓解了数据的噪声特性。

这两个问题导致在 UDT 等数据库中做出了特定的设计决策，其中某些没有跨语言对应关系的标签被归入其他类别。为了进行比较，有必要确保待比较的对象共享标签，这导致了多个层次的注释，使得语言特定的词类与更普遍的类别一起被标注。虽然这种方法仍然存在问题（例如，不同语言之间对语言特定词类的定义相差很大），但它确实允许在一个比标记的 PBC 所报道的更细粒度的层次上比较类型学特征，特别是通过努力识别语言中词之间的依赖关系。

本文报告了对标记 PBC [7] 的一次更新，试图通过将依赖标注从英文扩展到数据集中的每种语言，并将每个数据集转换为 CoNLL-U 格式来解决其中的一些缺点。[8] 这里的基本想法是，由于这些是平行语料库，依赖关系在对应句子/诗句的翻译等价物中理应保持一致。以下各部分介绍了用于标注依赖的方法（第 2 节），并显示从这些语料库中得出的量度与跨语言参考数据库中的及物子句基本词序判断相对应（第 3 节）。最后，我总结了一些总体观察结果，并提出了一些关于改进此数据集标注以进行额外的跨语言调查的可能方向（第 4 节）。

## 2 方法

由于将词性标注从高资源语言转移到低资源语言的效果在诸如名词和动词等常见类中相对较好（参见 Ring 2025a 及相关引用），我使用类似的方法将依赖信息转移到平行句子中。使用相同的语言专用分词器，我使用由 Ring (2025a) 训练的 IBM Model 2 词对齐模型来识别平行诗句中的翻译等价。由于 SpaCy 也有英文解析器，我然后将 SpaCy 处理的英文（目标）句子/诗句中的依赖信息和形态信息转移到源语言的平行诗句中的对应词。这一过程的具体细节包含在相关代码中，链接库中（脚注 7 ）。

这个过程允许自动传递依赖关系，前提是确切识别翻译的对应项。例如，在马太福音 8:1 中的下一个诗句（表 1 ）中，可以看到爱尔兰语（ISO 639-3: gle）文本带有英语（ISO 639-3: eng）信息的注释。虽然有些转移的信息显然不准确，但尽管两种语言之间的词序不同（爱尔兰语：VSO，英语：SVO），动词和相关论元还是被准确识别，这一点令人惊讶。[9]

# sent_id = 40008001
# ref_id = Matthew 8:1
# eng_text = When he came down from the mountain , great multitudes followed him .
# gle_text = Tháinig sé anuas ón sliabh agus lean sluaite móra é .

| id | word | pos | deprel | gloss_eng |
|---|---|---|---|---|
| 1 | Tháinig | VERB | advcl | gloss=came |
| 2 | sé | unk | _ | gloss=He |
| 3 | anuas | ADP | prep | gloss=down |
| 4 | ón | unk | _ | gloss=instantly |
| 5 | sliabh | NOUN | pobj | gloss=mountainside |
| 6 | agus | unk | _ | gloss=and |
| 7 | lean | VERB | root | gloss=followed |
| 8 | sluaite | NOUN | nsubj | gloss=crowds |
| 9 | móra | ADJ | amod | gloss=Large |
| 10 | é | unk | _ | gloss=it |
| 11 | . | PUNCT | punct | |

Table 1: 通过英语 标记依存关系的爱尔兰诗句

这个例子首先表明，通过平行语料库在语言之间传递这种依赖信息是可能的。同时，它也显示出这种自动化传递会引入相当多的噪音，某些语言的语料库可能比其他语言的语料库包含更多的噪音，这取决于其在各种语言特性上与英语的贴合程度。与之前关于原始 taggedPBC 数据集中噪音的观察一样，仍然有可能观测到跨语言的趋势。

[7] https://github.com/lingdoc/taggedPBC
[8] https://universaldependencies.org/format.html
[9] 由于篇幅限制，这里不包含语料库中存在的额外注释。

正如 Ring (2025a) 所指出的，计算类型学的研究人员有责任证明所使用的各种测量方法的有效性以及基础数据的有效性。在词性标注的情况下，Ring (2025a) 通过与最先进的标注器和手工标注的语料库进行比较，验证从英语转移过来的词性标签的准确性，涉及大量的（非英语）语言。在依存关系的情况下，验证转移依存关系的准确性存在一些问题。由于依存关系在某种意义上不像词类那样“固定”，同一个词可能在不同的从句中出现在多个依存关系中，因此更难评估其准确性。这将需要一个注释了依存关系的并行数据集，或者需要对当前数据集的一部分进行手工注释以获取依存关系。另一种可能性是对自动解析器进行评估，比如由 Üstün et al. (2020); Choudhary & O'riordan (2023) 开发的解析器。

然而，还有另一种可能性是评估通过依赖关系得出的各种测量结果与专家类型学确定结果的一致性。在这种意义上，我们略过狭隘的验证，转而选择具有已知/预期结果的更广泛的验证。我们将在接下来的部分中探讨从此数据集中提取的测量结果与现有专家分类之间的相关性。

## 3 与及物语序 的对应关系

标记的 PBC 的 CoNLL-U 格式版本包含了相当多的附加信息，而原始的词性标注数据集没有这些信息。虽然原始的标注数据集能够根据 N1 比率展示 SV、VS 和自由词序语言之间的区别，但该研究主要集中在根据与不及物动词相关的词序来区分语言，不及物动词自身只有一个（“主语”）论元。相反，转入 CoNLL-U 格式版本的依赖关系额外编码了论元的“主语”和“宾语”信息，从而允许在及物动词的句子/从句结构上进行六种区别（VSO、VOS、SVO、OVS、SOV、OSV）。

需要注意的是，所有语言中“主语”和“宾语”关系的存在在某种程度上是有争议的。这是因为“主语”通常指的是某语言中及物动词/从句的论元和不及物动词/从句的论元之间的特定对齐关系。在主宾结构语言中，“主语”被认为是及物从句的论元，其标记（以某种方式）与不及物从句的单一论元相似。这通常与动词事件的“施事”相对应（例如“他打她”［ tr ］与“他跑”［ intr ］）。相比之下，在作通格语言中，不及物从句的单一论元与及物从句的“承受者”或“接受者”论元（“宾语”）标记相同。然而，由于 UD 框架不在跨语言层面上区分这一点（而是留给特定语言的注释来澄清），我采用“主语”和“宾语”术语，并期待未来的注释工作能够解决这一重要问题。

尽管在注释中可能存在潜在的歧义和错误，我们仍然可以遵循 Ring (2025a) 的过程来推导 N1 比率，试图推导一种基于语料库的及物词序测量。为此，对于每种语言，我们只需计算其中同时标识“主语”和“宾语”依赖关系，并且有一个词被标记为“动词”的诗节数。对于六种及物词序模式中的每一种，我们可以通过用该模式出现的次数除以语料库中包含所有三个元素的句子/诗节的总数来推导出一个比率（或百分比/比例），从而实现 Levshina et al. (2023) 推荐的渐变观察。作为一个例子，图 1 绘制了具有不同词序的三种语言的词序比例：爱尔兰语（VSO; gle）、英语（SVO; eng）和印地语（SOV; hin）。

从语料库中可以获得这样的梯度观察是有趣的，但真正的问题在于它们是否有意义。为测试这一点，我从现有的类型学数据库（WALS、Grambank、Autotyp）中提取了及物词序信息，并观察了这些专家判断与从 CoNLL-U 标注的 PBC 语料库中得到的百分比之间的相关性。然而，这些数据库中的观察结果在可比性方面存在问题。例如，尽管 WALS 将个别语言识别为具备六种基本及物词序之一（或“自由”词序），Grambank 仅识别出三种动词位置之一（动词在首、动词在中、动词在末）的证据，以及词序是否“固定”或“自

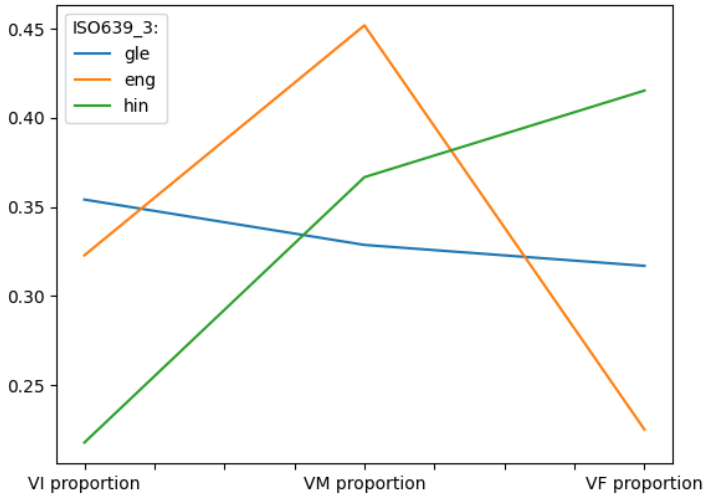


Figure 1: 三种语言的词序比例

由”。此外，虽然 Autotyp 类似于 WALS 识别基本及物词序（用“A”代替“S”），但某些语言是根据动词相对于其他元素的位置来识别的（例如，Tohono O’odham [ ood ] 被列为“Vxx”）。因此，为了符合可比性的需要，我将七种及物词序的可能性减少为四种：动词首位（“VI”，包括 VSO 和 VOS）、动词中位（“VM”，包括 SVO 和 OVS）、动词末位（“VF”，包括 SOV 和 OSV），以及“自由”。

将这三个类型数据库结合起来，使我们能够对比 961 种语言的词序分类，这些语言也出现在 taggedPBC 中。在图 2 - 4 中，我们可以观察到一些关于文集中每种观察到的词序模式比例的总体趋势，这些比例是根据其类型学分类得出的。有趣的是，单向 ANOVA（查看链接存储库中的统计数据，见注 7 ）显示，某些词序比例在某些分类之间具有显著的区分性，而在其他分类之间则没有。特别是，“自由”词序的语言与被分类为其他基本词序的语言极难区分。这是合理的，因为“自由”词序语言允许六种确认的模式中的任何变体，这意味着它们的一些语料库会显示出与 VI 语言一致的比例，而其他的则会显示出与 VM 或 VF 语言一致的比例。

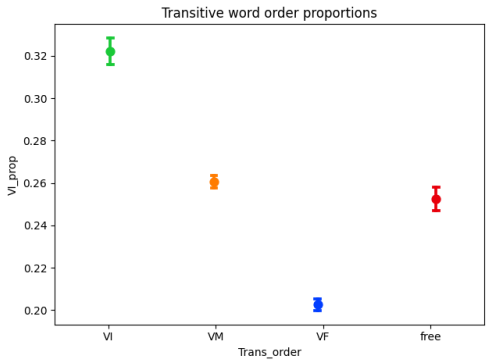


Figure 2: Verb-initial proportions

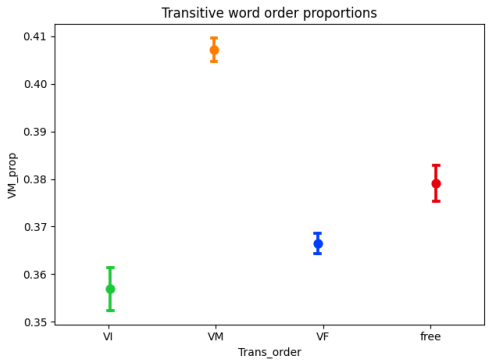


Figure 3: Verb-medial proportions

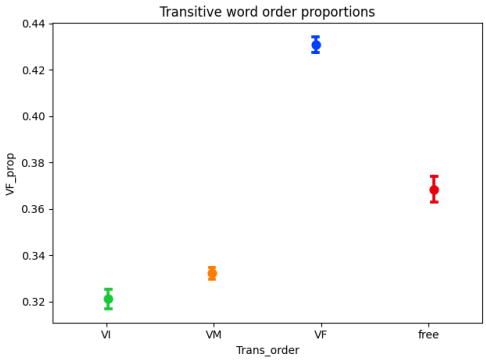


Figure 4: Verb-final proportions

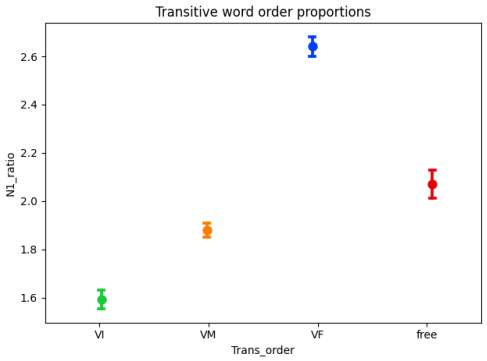


Figure 5: N1 ratio correspondences

不幸的是，这意味着仅仅根据语料库中词序出现的比例来确定语言是“固定”还是“自由”词序将非常困难。一旦将“自由”词序语言与“固定”词序语言区分开来，那么似乎就有可能将“固定”语言归类到剩下的三个类别之一。然而，从一开始就仅凭比例来识别哪些是“自由”哪些是“固定”似乎是不可能的，这使得这种努力显得徒劳。

然而，探索性分析揭示了另一种相关性。进行方差分析（ANOVA），以 N1 比率为因变量，以 VI、VM、VF 和自由分类为固定因素，显示 N1 比率明显区分了这些组（$p < 0.003$；见图 5 ）。令人惊讶的是，这一指标仅仅通过将以论元开头的诗句数量除以以谓词开头的诗句数量得出，使我们能够判断一种语言是否可以被归类为具有 VI、VM、VF 或自由词序，并与现有类型学数据库显示出明显的一致性。

## 4 结论

前面的章节显示，更新版的 TaggedPBC 将依存信息从英语转移，且与现有的（及物）词序分类相当吻合。虽然这个数据集仍然相当嘈杂，但它表明借助自动注释创建的大型数据集可以支持对粗略特征（如词序）的跨语言研究。有趣的是，通过跨语言转移，数据集中的所有语言都包含“主语”、“宾语”和“动词”的句子/诗句。此外，所有语言至少在六种及物词序模式中，每种模式都有一句/诗句被标记。

这里展示的词序研究不仅强调了梯度方法的重要性，还展示了识别离散词序类别的可能性。语料库的观察表明，所有语言在不同程度上展示了所有词序。同时，N1 比率与类型学数据库中专家判断之间的显著相关性突出了这些类别的有效性。有些语言明显优先或语法化一种顺序，并且这在语料库统计中有所反映。

然而，对于这种类型的数据库，还有许多问题需要解决。尽管这里报告的调查结果表明当前的数据集足以进行粗粒度分析，但当前的方法引入了一定程度的噪音。例如，词性标注和依存关系都是以英语为中心的，一些特定语言的标签（如数字分类词）在数据集中大多或完全缺失。此外，由于所用方法依赖词的对齐而非形态对齐，这个数据集中的语言都没有形态注释。在这里，我们要依靠个别语言的分词惯例通过空格进行词语分割。另外，尽管已经尝试识别特定词的 HEAD 关系，[10] 这一关系很难在翻译的基础上自动识别，这意味着从属从句不能与主句区分开，而这种问题可能会对识别诸如词序模式这类内容产生很大影响。

这些问题中的每一个都可以通过手动标注以及新方法来解决，或许是逐步地和渐进地。考虑一些可能性，通过利用手动标记的“黄金标准”语料库来训练更好的标注器，很有可

[10]在 UD CoNLL-U 格式中使用的意义上，字段 # 7 下：https://universaldependencies.org/format.html

能改善特定语言的现有词性标注，并且可以通过合并来自其他体裁的手动标记数据来扩展数据集。此外，即使特定语言的专家进行的手动标注不仅在词性标注而且在依存关系方面会有很大改进，当前数据集作为平行语料库的性质也适合于非专家标注者的支持。由于依存信息在平行的诗句/句子中预计会对齐，如果标注人员可以识别出特定单词的翻译等价物，那么相关的依存关系也可以被标注（具有高度的确定性），而不需要具体语言的详细知识。

在新方法方面，一个例子是通过对代表类型学特征多样性和类型学数据库中识别的词类的选定语言组进行人工标注的方式训练“通用标注器”的可能性。一旦为世界语言识别出完整的词类集，对相关语言数据集的一部分进行人工标注，可以观察与特定标签和相关词类分布相关的模式，类似于现有的跨语言标签转移方法（即 Imani et al. 2022 ）。具有与语言特定标签相关的统计先验可以使计算模型在实际标签未知的情况下，在一次性情况下预测或识别特定语言（如“数词分类器”）的标签。类似的改进可以应用于现有的多语言依赖关系标注方法（Choudhary & O’riordan 2023 ）。

总体而言，这个数据集代表了计算类属学（以及其他类型的跨语言研究）可以如何开展的重大变革。尽管对数据本身有许多潜在的改进，但我们能够通过许多语言观察粗粒度趋势这一长期目标已经值得重视。在计算工具和统计方法的辅助下，通过对这种并行数据的比较，我们可以学到很多关于人类语言的知识。尽管随着基础数据的进一步开发，这些发现可能略有变化，但我在这里展示了数据中发现的特征与世界语言中及物子句的词序观察之间具有相当好的相关性，这支持了语料库衍生特征在跨语言研究中是有价值的资产的观点。

## References

Baylor, Emi, Esther Ploeger, & Johannes Bjerva. 2023. The past, present, and future of typological databases in NLP. In Bouamor, Houda, Juan Pino, & Kalika Bali (eds.), Findings of the Association for Computational Linguistics: EMNLP 2023 , pp. 1163–1169. Singapore: Association for Computational Linguistics. URL https://aclanthology.org/2023.findings-emnlp.82/.

Baylor, Emi, Esther Ploeger, & Johannes Bjerva. 2024. Multilingual gradient word-order typology from universal dependencies. URL https://arxiv.org/abs/2402.01513. 2402.01513.

Bickel, Balthasar, Johanna Nichols, Taras Zakharko, Alena Witzlack-Makarevich, Kristine Hildebrandt, Michael Rießler, Lennart Bierkandt, Fernando Zúñiga, & John B Lowe. 2023. The AUTOTYP database (v1.1.1).

Bjerva, Johannes. 2024. The role of typological feature prediction in nlp and linguistics. Computational Linguistics , 50(2):781–794. ISSN 0891-2017. URL https://doi.org/10.1162/coli_a_00498. https://direct.mit.edu/coli/article-pdf/50/2/781/2457439/coli_a_00498.pdf.

Choudhary, Chinmay & Colm O’riordan. 2023. Multilingual end-to-end dependency parsing with linguistic typology knowledge. In Beinborn, Lisa, Koustava Goswami, Saliha Muradoğlu, Alexey Sorokin, Ritesh Kumar, Andreas Shcherbakov, Edoardo M. Ponti, Ryan Cotterell, & Ekaterina Vylomova (eds.), Proceedings of the 5th Workshop on Research in Computational

Linguistic Typology and Multilingual NLP , pp. 12–21. Dubrovnik, Croatia: Association for Computational Linguistics. URL https://aclanthology.org/2023.sigtyp-1.2/.

Dryer, Matthew S. & Martin Haspelmath (eds.). 2013. WALS Online (v2020.4) . Zenodo.

Hammarström, Harald, Robert Forkel, Martin Haspelmath, & Sebastian Bank. 2024. Glottolog 5.1 . Leipzig: Max Planck Institute for Evolutionary Anthropology. URL http://glottolog.org. Accessed on 2025-04-03.

Imani, Ayyoob, Silvia Severini, Masoud Jalili Sabet, François Yvon, & Hinrich Schütze. 2022. Graph-based multilingual label propagation for low-resource part-of-speech tagging. In Goldberg, Yoav, Zornitsa Kozareva, & Yue Zhang (eds.), Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing , pp. 1577–1589. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.

Levshina, Natalia, Savithry Namboodiripad, Marc Allassonnière-Tang, Mathew Kramer, Luigi Talamo, Annemarie Verkerk, Sasha Wilmoth, Gabriela Garrido Rodriguez, Timothy Michael Gupton, Evan Kidd, Zoey Liu, Chiara Naccarato, Rachel Nordlinger, Anastasia Panova, & Natalia Stoynova. 2023. Why we need a gradient approach to word order. Linguistics , 61(4):825–883. URL https://doi.org/10.1515/ling-2021-0098.

List, Johann-Mattis, Robert Forkel, Simon J. Greenhill, Christoph Rzymski, Johannes Englisch, & Russell D. Gray. 2022. Lexibank, a public repository of standardized wordlists with computed phonological and lexical features. Scientific Data , 9(1):316. URL https://doi.org/10.1038/s41597-022-01432-0.

Mayer, Thomas & Michael Cysouw. 2014. Creating a massively parallel Bible corpus. In Proceedings of The International Conference on Language Resources and Evaluation (LREC) , pp. 3158–3163. Reykjavik.

Ring, Hiram. 2025a. The taggedPBC : Annotating a massive parallel corpus for crosslinguistic investigations. URL https://arxiv.org/abs/2505.12560. 2505.12560.

Ring, Hiram. 2025b. Word length predicts word order: "min-max"-ing drives language evolution. URL https://arxiv.org/abs/2505.13913. 2505.13913.

Skirgård, Hedvig, Hannah J. Haynie, Damián E. Blasi, Harald Hammarström, & et al. 2023. Grambank reveals global patterns in the structural diversity of the world's languages. Science Advances , 9.

Üstün, Ahmet, Arianna Bisazza, Gosse Bouma, & Gertjan van Noord. 2020. UDapter: Language adaptation for truly Universal Dependency parsing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP) , pp. 2302–2315. Online: Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.emnlp-main.180.

Zeman, Daniel, Joakim Nivre, Mitchell Abrams, & et al. 2024. Universal dependencies 2.14. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.