通用依存树库是否越来越一致？EnglishUD 成绩报告

Amir Zeldes乔治城大学
amir.zeldes@georgetown.edu

Nathan Schneider乔治城大学
nathan.schneider@georgetown.edu

arXiv:2302.00636v1 [cs.CL] 1 Feb 2023

摘要

通用依存项目近期在整合指南和树库方面的努力，使人们期待：对于英语等拥有多个语料库的高资源语言，联合训练和数据集比较正变得越来越可行。本文聚焦于两个最大的 UD 英语树库，考察数据整合的进展，并回答若干问题：UD 英语树库内部是否变得更加一致？它们是否变得更加相似？相似程度如何？联合训练是否可取？如果可取，从哪个 UD 版本开始适用？我们的结果表明，尽管整合已取得进展，联合模型仍可能受到不一致问题的影响，进而妨碍其利用更大规模训练数据的能力。

1 引言

通用依存项目[1]（de Marneffe 等人，2021年）在过去几年间不断发展，如今涵盖了 100 多种语言以及 200 多个树库，这意味着现在有多种语言拥有多个树库，并带有丰富的形态句法及其他标注。对于英语等资源丰富的语言，多个树库尤为常见；目前英语的数据分布在 9 个不同的代码库中，词元总数超过 762,000 个（截至 UD v2.11）。资源如此丰富当然是件好事，但这也引发了关于同一种语言的多个通用依存树库之间一致性的问题，既包括标注指南的理论问题，也包括联合训练多个数据集对依存句法分析及其他自然语言处理应用的实际价值问题。

本文聚焦于最大的两个英语通用依存树库：英语网络树库（EWT，Silveira 等，2014年）和乔治城大学多层语料库（GUM，Zeldes，2017）。[2] 尽管这两个数据集都旨在遵循通用依存指南，但它们的起源截然不同：EWT 是从较早的成分树库（Bies 等，2012）转换为斯坦福依存关系（de Marneffe 等，2006），随后又转换为通用依存；GUM 则一直采用斯坦福依存关系进行原生标注，直到 2018 年才转换为通用依存（Peng 和 Zeldes，2018），之后又通过原生通用依存标注逐步加入更多材料。考虑到指南不断修订和澄清，数据集之间可能存在系统性差异；通用依存维护者（包括作者）一直致力于在不同通用依存版本中整合这些差异。

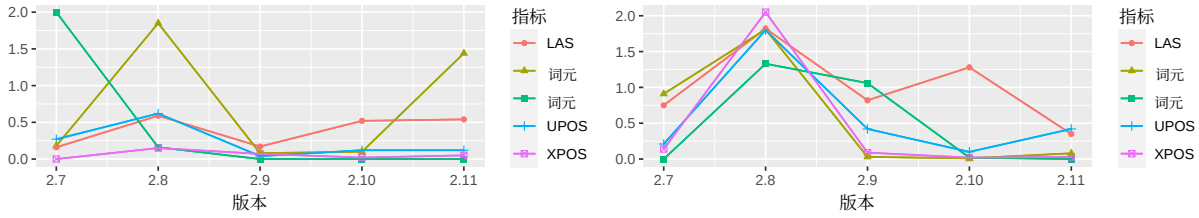
尽管存在潜在问题，自然语言处理工具正日益将通用依存数据集合并起来进行联合训练：例如，斯坦福大学广受欢迎的 Stanza 工具包（Qi 等，2020年）默认使用名为 `combined`的模型进行英语标注和句法分析，该模型在 EWT 和 GUM 上训练（包括 GUM 中的 Reddit 子集）。[3] 因此，我们认为现在正是探讨这一问题的合适时机：即使是规模最大、开发最活跃的英语通用依存树库，是否真的彼此兼容？如果不兼容，差异有多大？它们是否正逐渐趋同，还是随着版本更新而渐行渐远？无论这些问题的答案如何，在 EWT 和 GUM 上联合训练是否是个好主意？如果是，考虑到数据不断修订，应从哪个通用依存版本开始？

2 相关工作

此前许多关于通用依存一致性的研究都聚焦于跨语言比较，尤其是寻找可能的错误。一些论文采用“广度优先”的自动化方法来识别各种不一致之处（de Marneffe 等人，2017），但需要注意的是，许多类型的差异很难检测。另一些研究则更聚焦于特定现象，

[1] https://universaldependencies.org [2] 由于许可限制，GUM Reddit 数据（Behzad 和 Zeldes，2020年）有单独的代码库，但我们会在下文合并这两个代码库。

[3] 虽然本文聚焦于英语，但其他拥有多个数据集的通用依存语言也存在同样的情况。



(a) EWT **(b)** GUM

图 1： EWT 和 GUM 各标注层随版本变化的情况。纵轴数值为百分比 es.

例如，Bouma 等人（2018年）使用通用依存 v2.1 表明，`expl` 关系在不同语言中对可比案例的使用方式有所不同。Sanguinetti 等人（2022年）展示了在 v2.6 中，通用依存语言对来自网络的用户生成内容进行标注时所采用的广泛做法。Dönicke 等人（2020年）也使用通用依存 v2.5 展示了通用依存语言内部存在的不一致，其中发现，通用依存中标注不一致程度最高的前 20 个关系中，有两个来自英语。在 7 个数据集中，`compound` 和 `csubj` 的表现不同（其中只有后者在 EWT 和 GUM 中存在显著差异，不过作者指出，GUM“可能只是包含了更多带有虚词的句子”）。Aggarwal 和 Zeman（2020年）考察了通用依存 v2.5 中词性（POS）标签的一致性，发现大多数语言内部的词性标注相对一致。

针对跨语料库句法分析准确率的研究较少（Alonso 和 Zeman 2016 研究了通用依存 v1.3 上的西班牙语，Droganova 等人 2018 研究了使用通用依存 v2.2 的俄语），而考察句法分析一致性和稳定性的研究就更少了（Kalpakchi 和 Boye，2021年）。然而，据我们所知，此前没有研究考察不同通用依存版本间一致性的变化，即树库间的兼容性是否随着时间推移而提升、提升幅度有多大，以及哪些标注的兼容性有所提升?

3 数据发生了哪些变化?

为了了解两个语料库的数据在不同版本间发生了怎样的变化，我们使用官方的 CoNLL 2018 UD 共享任务（Zeman 等人，2018年）评分器，将每个语料库过去六个版本中的每个版本与其后续版本进行比较，并将更新后的版本视为改进后的“金标准”。[4] 由此可得出各项通用依存指标的分数，例如带标签依存准确率（LAS）、通用词性（UPOS）和英语专用词性（XPOS），以及词元化和分词。图1 展示了各数据集每对版本之间的分数差异

[4]由于早于 GUM 转换为通用依存的时间，因此无法进行更早的比较。

下文将针对各语料库分别讨论。举例来说，以 EWT 的 v2.7 作为 v2.6 的修正版，我们发现分词错误率为 2%（绿色部分），表明分词发生了显著变化；而到 v2.8 时变化不到 0.2%，升级到 v2.9 时分词则毫无变化。

需要注意的是，跨版本处理数据时，GUM 与 EWT 不同，其内容并非固定不变：语料库每年都会新增材料。因此，在下面的概述中，我们将评估范围固定为仅包含自 v2.6 起就已存在的文档（136 篇文档，12 万词元）。在 §4 中，我们将探讨同时使用这一固定子集和整个语料库的情形（在 v2.11 中共有 193 篇文档，18 万词元）。[5]

3.1 EWT

下面我们解释连续版本之间差异较大的主要原因。

分词 在 v2.7 中，为大多数附着词（例如 *'ll*）和缩约形式（*don't*）添加了多词词元（MWTs），v2.8 中则补充了一些遗漏项。此后的版本基本没有对分词进行任何更改。

标注 v2.7 中出现了中等程度的通用词性标签变化（许多 WH词被改为 `SCONJ`），v2.8 中也有变化（专有名称中的形容词和动词改为 `ADJ` 和 `VERB`，此前为 `PROPN`，与扩展词性标签 `NNP`保持一致）；GUM 后来也采用了这一更改，见下文。扩展词性标签的变化较小，在 v2.8 中达到峰值，涉及*of course*、*at least*和美国等特定表达。

词元化 词元错误已全面修正，但 v2.8 中词元变更的主要来源包括专有名称中实词词元的大小写调整、代词*I*的词元调整，以及移除比较级或最高级

[5]随后发布的规模更大的 GUM v9，包含 203K 个词元和 213 份文档，发布时间与本文最终稿截止日期大致相同，因此未能及时进行评估。

形式的词元。在 v2.11 中，制定了一项有关物主代词词元的新政策，以消除与 GUM 的一项关键差异。

依存关系如图1a所示，LAS 最大的变化发生在 2.8 版本，当时新标注的 ADJ 名称中的词元触发了 amod；2.10 版本将*X, so Y* 结构的分析改为 parataxis （以及其他变化）；2.11 版本则调整了嵌套主语（nsubj:outer）、关系结构和分裂句。

3.2 GUM

与 EWT 类似，GUM 在各层面上已变得更加稳定，自 v2.9 以来，扩展词性标签或词元均几乎没有变化。不过，早期版本中有几项重大修订。许多变化同样只是错误修正所致，但一些系统性变化包括以下内容。

分词在 v2.8 中发生了重大变化，引入多词词元以匹配 EWT 的变化。v2.9 中的其他重大变化则源于调整词语分词方式，使其与 EWT 及其他近期 LDC语料库保持一致；这些语料库会对带连字符的复合词进行分词（例如，v2.7 将 *data-driven* 作为一个词元，而 v2.8 则将其分为三个词元，与 EWT 一样）。

标注 在 v2.8 中也出现了类似变化，这是因为引入了用于连字符的 HYPH 标签，例如复合词“data-driven”中的连字符；此外，还移除了方括号专用的扩展词性标签（左/右方括号的 -LSB- 和 -RSB- 被合并到圆括号标签 -LRB-/-RRB-中，这再次与 EWT 保持一致）。相比之下，通用词性标签的变化更为显著，主要是因为专有名称中的动词/形容词发生了变化，如上文 EWT 所示。之后 v2.9 和 2.11 中通用词性标签的变化，则源于部分代词限定词（扩展词性标签 DT）被重新标注为 DET 而非 PRON（*some*、*all*、*both*），以及 WH 从属连词分别从 SCONJ 改为 ADV；这些变化同样与 EWT 的变化一致。

词元化 在很大程度上反映了连字符用法的变化（例如，“data-driven”在 v2.8 中不再是一个词元），以及名称中从 PROPN 改为 VERB或 ADJ 的变化，因为“*Glowing*”在“*the GlowingSea*”中的词元，从“*Glowing*”（基于其为 PROPN）改成了“*Glow*”（作为一个 VERB）。

依存关系 这里，过渡到新的分词和标注名称也导致 v2.8 中出现变化；但我们还看到 v2.10 达到峰值，这主要是由于专有名称依存关系的整合（将 flat 改为句法透明的分析），对省略的识别更加积极（包括提升论元）和 orphan 关系，以及移除了一些 dep 关系的用法。

4 跨语料库句法分析

跨语料库结果 为了检验 EWT 和 GUM 的内部一致性及彼此之间的一致性是否有所提高，我们在每个语料库的各个版本上训练句法分析器，并在两个语料库上对其进行测试。如果每个语料库的一致性都在提高，我们预计各版本的得分都会更高；如果跨语料库模型的得分不断上升，我们便推断不同语料库之间的数据一致性正在提高。为确保比较公平，我们将 GUM 的训练集和测试数据固定为从 v2.6 起便已提供的那些文档。

表格中的结果1 表明，语料库内部的得分确实随着每个版本的更新略有提高（所有得分均为 3 次运行的平均值，使用 Diaparser，Attardi 等（2021）提出的最新基于 Transformer 的双仿射依存句法分析器）。跨语料库得分明显较低，但也在提高：在 v2.6 中，EWT 域内 LAS 为 90.24；到了 v2.11，略有提升至 90.9；而在 GUM 上训练、在 EWT 上测试的得分则从 81.78 升至 84.27。反过来，GUM 域内得分从 LAS=87.9 提升至 89.48；对于在 EWT 上训练的句法分析器，得分则从 83.89 升至 84.74。两个语料库的宏平均也呈现稳步上升趋势，GUM 上的增幅更明显。在所有情况下，v2.11 在各项指标上都是迄今为止表现最好的版本。

然而，由于这些实验仅限于 UD GUM v2.6 中较小的文档子集，因此并不能反映当前的自然语言处理工具（这些工具在当前 UD 代码库中的所有文档上进行训练），也无法说明联合训练是否可取。

联合训练结果 在这一系列实验中，我们对两个语料库进行联合训练，并比较两种场景：SUBSET 场景将 GUM 训练数据限制为 v2.6 子集，而 ALL 则在每个版本中使用所有可用的 GUM 文档进行训练；为确保公平，评分始终仅针对 v2.6 测试集中的文档，而该测试集是后续所有版本测试集的子集。[6]

表2 同样表明，这里也只有随着时间推移而不断改进。使用 GUM 的全部文档优于仅使用 GUM 子集，但实际上会使 EWT 的表现略有下降，这大概是因为纳入了更多域外数据

[6]请注意，GUM 在 v2.7 中没有新增文档，因此直到 v2.8，SUBSET 和 ALL 的评分都相同。

训练集	版本	EWT 测试集		GUM 测试集		宏平均	
		无标注依存准确率（标准差）	LAS（标准差）	无标注依存准确率（标准差）	LAS（标准差）	无标注依存准确率（标准差）	LAS（标准差）
EWT	v2.6	92.82 0.132	90.24 0.066	87.81 0.073	83.89 0.023	90.31 0.059	87.07 0.025
	v2.7	92.84 0.037	90.25 0.173	87.87 0.088	84.19 0.074	90.35 0.062	87.22 0.114
	v2.8	92.93 0.060	90.42 0.090	87.97 0.078	84.90 0.028	90.45 0.065	87.66 0.042
	v2.9	92.88 0.107	90.41 0.131	87.57 0.148	84.36 0.105	90.23 0.098	87.38 0.117
	v2.10	93.06 0.082	90.70 0.158	87.81 0.084	84.72 0.138	90.44 0.082	87.71 0.088
	v2.11	**93.18** 0.142	**90.90** 0.139	**88.05** 0.260	**84.74** 0.289	**90.62** 0.196	**87.82** 0.207
GUM	v2.6	86.53 0.357	81.78 0.397	91.37 0.201	87.90 0.141	88.95 0.187	84.84 0.209
	v2.7	86.69 0.336	82.28 0.322	91.66 0.156	88.24 0.284	89.18 0.242	85.26 0.299
	v2.8	87.02 0.133	82.90 0.214	91.88 0.132	88.86 0.159	89.45 0.002	85.88 0.041
	v2.9	87.42 0.143	83.43 0.025	91.88 0.300	88.78 0.281	89.65 0.219	86.11 0.140
	v2.10	87.53 0.190	83.79 0.191	92.16 0.216	89.24 0.191	89.85 0.203	86.51 0.191
	v2.11	**88.23** 0.198	**84.27** 0.095	**92.28** 0.137	**89.48** 0.224	**90.26** 0.121	**86.88** 0.132

表1： 跨语料库句法分析得分（三次运行平均值及标准差）

训练集	版本	EWT 测试集		GUM 测试集		宏平均	
		无标注依存准确率（sd）	LAS（sd）	无标注依存准确率（sd）	LAS（sd）	无标注依存准确率（sd）	LAS（sd）
联合子集	v2.6	92.38 0.044	89.59 0.108	90.08 0.366	86.80 0.326	91.23 0.177	88.20 0.146
	v2.7	92.31 0.078	89.61 0.072	90.15 0.311	86.96 0.360	91.23 0.122	88.29 0.148
	v2.8	92.49 0.159	89.99 0.128	90.51 0.351	87.86 0.449	91.50 0.154	88.92 0.195
	v2.9	92.39 0.324	89.80 0.278	90.63 0.392	87.91 0.415	91.51 0.086	88.85 0.114
	v2.10	92.62 0.034	90.24 0.058	90.51 0.418	87.86 0.381	91.56 0.192	89.05 0.163
	v2.11	**92.92** 0.072	**90.58** 0.052	**90.75** 0.073	**87.94** 0.059	**91.83** 0.064	**89.26** 0.045
联合全部	v2.6	92.38 0.044	89.59 0.108	90.08 0.366	86.80 0.326	91.23 0.177	88.20 0.146
	v2.7	92.31 0.078	89.61 0.072	90.15 0.311	86.96 0.360	91.23 0.122	88.29 0.148
	v2.8	92.07 0.277	89.55 0.312	91.26 0.267	88.72 0.247	91.66 0.077	89.14 0.066
	v2.9	92.27 0.154	89.77 0.287	90.81 0.084	88.12 0.123	91.54 0.110	88.95 0.176
	v2.10	92.18 0.018	89.86 0.010	91.54 0.170	88.99 0.211	91.86 0.092	89.43 0.110
	v2.11	**92.54** 0.259	**90.11** 0.240	**91.71** 0.426	**89.11** 0.534	**92.13** 0.147	**89.61** 0.181

表2： 联合训练句法分析分数（三次运行平均值及标准差）

从 EWT 的角度来看。尽管如此，JOINT*ALL*在 EWT 上的表现也随着时间推移而提升。

表1 中的最佳域内成绩始终优于最佳联合训练成绩，这表明新增数据并不能完全弥补各语料库中不同体裁带来的干扰，以及可能仍然存在的标注不一致。这并不令人意外，因为体裁对自然语言处理的表现至关重要（Zeldes 和 Simonson，2016年；Müller-Eberstein 等人，2021年）。事实上，在其他语言的通用依存句法分析中（参见希伯来语研究 Zeldes 等人，2022年；西班牙语研究 León，2020年），以及其他任务中（例如篇章分析，Peng 等人，2022年；Liu 和 Zeldes，2023年），也观察到了类似的权衡：数据量增加有益，但异质性增加有害。

不过，这一差距正在缩小：联合模型在 EWT 上提升了约 1 个百分点，仅落后于最佳域内模型 0.32 个百分点；在 v2.11 中的最佳 ALL 场景下，它在 GUM 上提升了 2.31 个百分点。或许更重要的是，宏平均可能更能反映“现实世界中的”句法分析器模型对任何未见过的体裁数据的适用性（因为宏观测试集包含最多的目标体裁），如今已达到LAS=89.61，距离每个语料库的最佳模型都不到一个百分点。

由于最佳联合结果也出现在 v2.11，因此我们似乎可以这样回答本文开头提出的问题：联合训练从未像现在这样值得采用；联合训练的表现始终紧随域内训练之后，但差距一直在缩小，如今已非常小；对于完全未见过的新数据，联合模型如今看来是个非常好的选择。联合 SUBSET 模型在 EWT 上紧随其后，而联合 ALL 模型在 GUM 上位居第二。不过，增加第二个语料库所带来的更多数据，表现仍不如单独进行域内训练，这表明语料库之间仍存在不一致之处，希望后续通用依存版本能够加以改进。

关于当前联合训练的句法分析器实际上在哪些方面出错，我们请读者参阅图2 中的

附录中的混淆矩阵。矩阵显示，尽管训练所用的数据集各不相同，两个测试集上最常见的错误始终是混淆 nmod 和obl，这通常对应于介词短语附着错误。其他系统性错误并不常见，且主要涉及值得注意的名称子类别及其他类型的术语。一种反复出现的子类型是：GUM 的dep 标签被误认为是 EWT 的 nummod ，用于非计数修饰语的数值修饰语（例如，“第3页”中的“3”在 GUM 中标为 dep ，但在 EWT 中标为 nummod ；GUM 仅在“3 页”这类计数情形中使用 nummod ）。Schneider 和 Zeldes（2021年）讨论了其中若干差异，这些差异也构成了进一步整合的目标。

另一个可能值得关注的问题是 compound 关系：GUM 训练的模型会将 EWT 中若干黄金标准关系预测为这些关系，而 EWT 模型也会将 GUM 中若干黄金标准关系预测为这些关系。这些现象似乎很可能是将 EWT 的黄金标准成分标注自动转换为依存关系时留下的残余产物；在转换过程中，若干复杂名词短语被分析为复合关系，例如斯里兰卡 或香港等名称（在 EWT 中为右中心 compound，在 GUM 中则为左中心 flat ），以及*cordon-blu*（原文如此）等借用的外来词或短语（同样在 EWT 中为右中心，在 GUM 中则会采用扁平关系）；还有一些复杂嵌套短语，在 EWT 中被分析为左分支结构，例如，*Mar-vel Consultants, Inc.* 以*Inc.* 为中心词，并在 EWT 中带有两个compound 依存项。在 GUM 中，中心词则会是*Consultants* ，而*Inc.* 作为 acl；对于词汇化情形则采用 flat（GUM 中有一例电影《怪物公司》）。类似地，带有 XPOS NNP 的首字母大写的形容词修饰语，有时在 EWT 中会被标注为 compound ，导致 GUM 训练的模型预测为 amod ，反之亦然（例如伊斯兰主义军官或复兴党破坏分子）。

5 讨论

本文回顾了整合规模最大的通用依存英语语料库 EWT 和 GUM 的进展。结果显示，数据正逐渐趋于一致：单语料库训练仍略胜联合训练一筹，但联合模型的表现已几乎同样出色，而且可能稳健得多。随着整合工作持续推进，我们希望看到联合模型超越域内训练，并且这种一致性进一步扩展到其他英语数据集和其他通用依存语言。

参考文献

Akshay Aggarwal 和 Daniel Zeman。2020年。估算一种语言中不同树库的词性标注一致性。载于第十九届树库与语言学理论国际研讨会论文集，第93–110 页，德国杜塞尔多夫。计算语言学协会。

Héctor Martínez Alonso 和 Daniel Zeman。2016年。AnCora树库的通用依存。自然语言处理，57:91–98。

Giuseppe Attardi、Daniele Sartiano 和 Maria Simi。2021年。用于增强型通用依存的双仿射依存关系与语义图句法分析。载于第*17*届国际解析技术会议及 *IWPT 2021* 增强型通用依存句法分析共享任务论文集（*IWPT 2021*），第184–188页，线上。计算语言学协会。

Shabnam Behzad 和 Amir Zeldes。2020年。一种面向稳健 Reddit 词性标注的跨体裁集成方法。载于第*12*届网络作为语料库研讨会论文集，第50–56页，法国马赛。欧洲语言资源协会。

AnnBies、JustinMott、Colin Warner 和Seth Kulick。2012。英语网络树库。技术报告 LDC2012T13，语言数据联盟，宾夕法尼亚州费城。

Gosse Bouma、Jan Hajic、Dag Haug、Joakim Nivre、Per Erik Solberg 和 Lilja Øvrelid。2018年。通用依存关系树库中的虚词。载于《第二届通用依存关系研讨会论文集》（*UDW 2018*），第18–26页，比利时布鲁塞尔。计算语言学协会。

Marie-Catherine de Marneffe、Matias Grioni、Jenna Kanerva 和 Filip Ginter。2017年。评估通用依存关系语料库的标注一致性。载于《第四届国际依存语言学会议论文集》（*Depling 2017*），第108–115页，意大利比萨。林雪平大学电子出版社。

Marie-Catherine de Marneffe、Bill MacCartney 和 Christopher D. Manning。2006年。从短语结构解析生成带类型的依存句法解析。载于《*LREC 2006*论文集》，第449–454页，意大利热那亚。

Marie-Catherine de Marneffe、Christopher D. Man- ning、JoakimNivre 和 Daniel Zeman。2021年。通用依存关系。 计算语言学，47(2):255–308。

Tillmann Dönicke、Xiang Yu 和 Jonas Kuhn。2020年。识别并处理 UD 中跨树库的不一致问题：一项试点研究。载于 《第四届通用依存关系研讨会论文集（*UDW 2020*）》，第 67–75 页，西班牙巴塞罗那（线上）。计算语言学协会。

Kira Droganova、Olga Lyashevskaya 和 Daniel Zeman。2018年。单语语料库的数据转换与一致性：俄语 UD 树库。载于《第十七届树库与语言学理论国际研讨会论文集（*TLT 2018*）》，第 53–66 页，瑞典林雪平。林雪平大学电子出版社。

Dmytro Kalpakchi 和 Johan Boye。2021年。微小的变化也会产生影响：通用依存句法分析器一致性的案例研究。载于第五届通用依存研讨会论文集（*UDW，SyntaxFest 2021*），第96–108页，保加利亚索非亚。计算语言学协会。

Fernando Sánchez León。2020年。结合不同的句法分析器和数据集进行 CAPITEL UD 句法分析。载于伊比利亚语言评测论坛论文集（*IberLEF2020*），与西班牙自然语言处理学会第*36*届会议（*SEPLN 2020*）同期举行，*2020*年*9*月*23*日，西班牙马拉加，第2664卷，*CEUR* 研讨会论文集，第39–44页。CEUR-WS.org。

Yang Janet Liu 和 Amir Zeldes。2023年。为什么话语句法分析无法泛化？对数据多样性影响的全面调查。载于*EACL 2023* 论文集，克罗地亚杜布罗夫尼克。

Max Müller-Eberstein、Rob van der Goot 和 Barbara Plank。2021年。体裁作为跨语言依存句法分析的弱监督。载于*2021*年自然语言处理实证方法会议论文集，第4786–4802页，线上及多米尼加共和国蓬塔卡纳。计算语言学协会。

Siyao Peng、Yang Janet Liu 和 Amir Zeldes。2022年。GCDT：用于多体裁和多语言话语分析的中文 RST 树库。载于 第二届计算语言学协会亚太分会会议暨第十二届自然语言处理国际联合会议（第二卷：短论文）论文集，第382–391页，仅在线发布。计算语言学协会。

Siyao Peng 和 Amir Zeldes。2018年。通往通用依存的道路殊途同归：通过多层标注将斯坦福和宾夕法尼亚句法分析结果转换为英语通用依存。载于语言标注、多词表达与构式联合研讨会论文集（*LAW-MWE-CxG-2018*），第167–177页，新墨西哥州圣菲。

Peng Qi、Yuhao Zhang、Yuhui Zhang、Jason Bolton 和 Christopher D. Manning。2020年。Stanza：面向多种人类语言的 Python 自然语言处理工具包。载于计算语言学协会第*58*届年会论文集：系统演示，第101–108页。

ManuelaSanguinetti、LaurenCassidy、Cristina Bosco、Özlem Çetino˘glu、Alessandra Teresa Cignarella、Teresa Lynn、Ines Rehbein、Josef Ruppenhofer、Djamé Seddah 和 Amir Zeldes。2022年。对用户生成内容进行树库标注：基于通用依存的指南、语料库和统一建议概述。语言资源与评测。

Nathan Schneider 和 Amir Zeldes。2021。通用依存中的特殊名词性结构。载于第五届通用依存关系研讨会论文集（*UDW，SyntaxFest 2021*），第160–172页，保加利亚索非亚。计算语言学协会。

Natalia Silveira、Timothy Dozat、Marie-Catherine de Marneffe、Samuel Bowman、Miriam Connor、John Bauer 和 Christopher D. Manning。2014年。英语金标准依存语料库。载于第九届语言资源与评测国际会议论文集（*LREC-2014*）。

Amir Zeldes。2017。GUM语料库：在课堂中创建多层资源。语言资源与评测，51(3)：581–612。

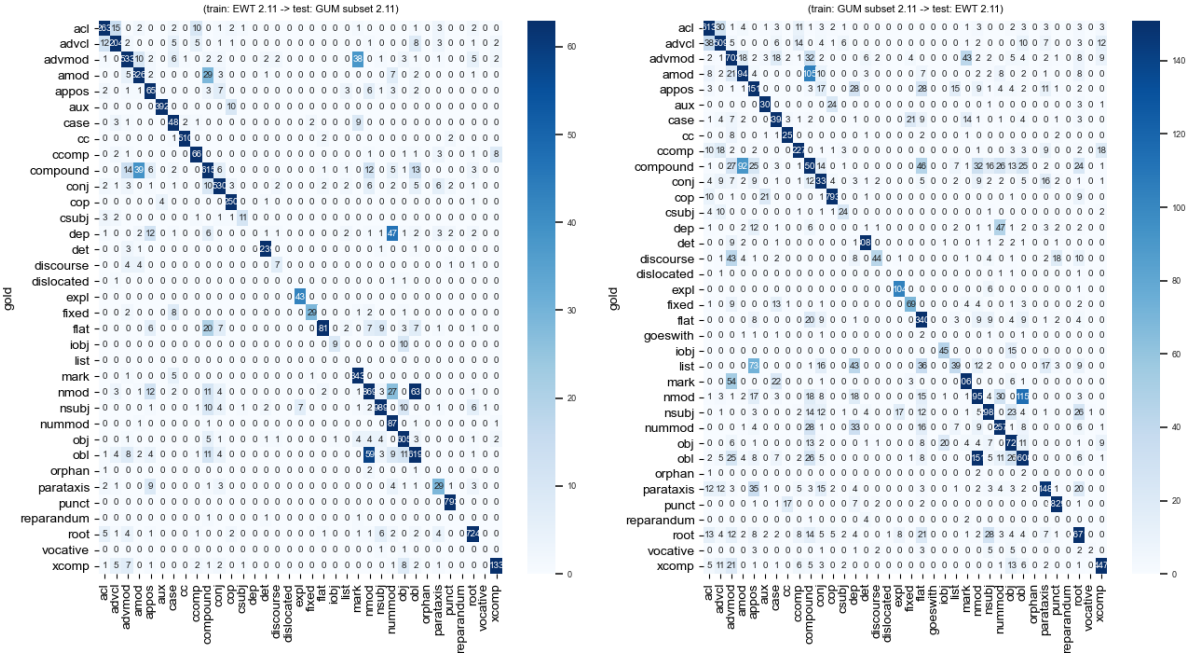
Amir Zeldes、Nick Howell、Noam Ordan 和 Yi- fat Ben Moshe。2022年。新一轮希伯来语通用依存树库构建与跨领域句法分析。载于*EMNLP 2022* 论文集，阿布扎比，阿联酋。

Amir Zeldes 和 Dan Simonson。2016年。GUM 的不同风味：评估体裁和句子类型对多层语料标注质量的影响。载于与 *ACL 2016*（*LAW-X 2016*）联合举办的第十届语言学标注研讨会论文集，第 68–78 页，柏林，德国。计算语言学协会。

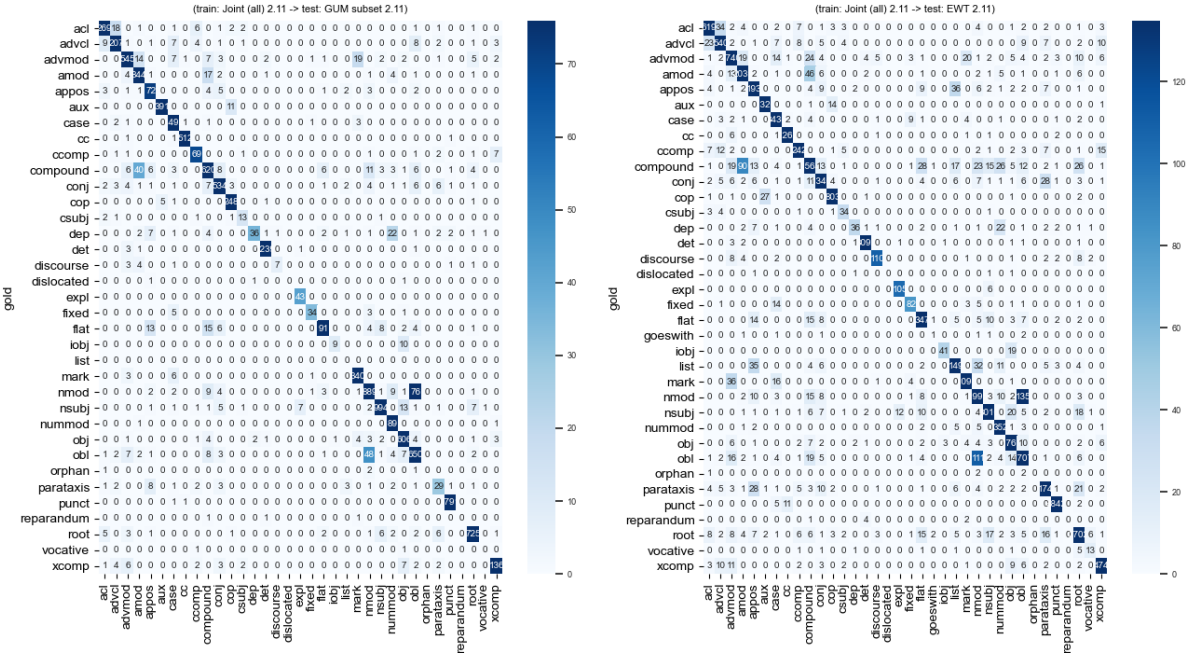
Daniel Zeman、Jan Haji˘c、Martin Popel、Martin Pot- thast、Milan Straka、Filip Ginter、Joakim Nivre 和 Slav Petrov。2018年。CoNLL 2018 共享任务：从原始文本到通用依存的多语言句法分析。载于*CoNLL 2018* 共享任务论文集：从原始文本到通用依存的多语言句法分析，第 1–21 页，布鲁塞尔，比利时。计算语言学协会。

A 混淆矩阵

图 2给出了联合场景和跨语料库场景下的依存关系预测混淆矩阵（不考虑附着正确与否），测试集分别为 GUM（左）和 EWT（右）。在所有情况下，最常被混淆的错误都是 `obl`和 `nmod`，且两个方向上都如此，这在很大程度上对应于介词短语附着歧义错误（即在“用叉子吃披萨”中介词短语高附着于动词，而在“吃一块带凤尾鱼的披萨”中低附着于宾语名词）。这些错误令人鼓舞，因为它们不太可能反映两个语料库之间的标注实践差异。



(a) 跨语料库训练的依存关系错误 g



(b)联合训练的依存关系错误

图 2： 跨语料库 (a) 和联合语料库 (b) 在两个测试集上的依存关系预测混淆矩阵；对 GUM 使用 GUM v2.6 文档子集，并采用每个实验中表现处于平均水平的句法分析器模型。