低资源语言技术语料库构建方法：南库尔德语和拉基语案例

SinaAhmadi♠ ZahraAzin♣ Sara Belelli♦ Antonios Anastasopoulos♠♣计算机科学系，乔治·梅森大学，美国弗吉尼亚州费尔法克斯♣卡尔顿大学语言学与语言研究学院，加拿大♦图西亚大学，维泰博，意大利{sahmad46,antonis}@gmu.edu,
taraazin@cmail.carleton.ca, sarabelelli@gmail.com

摘要

资源匮乏且濒危语言社群在语言技术领域面临的主要挑战之一，是语言数据的缺失或匮乏。南库尔德语和拉基语的南部变体也面临这一问题：可用资源非常有限，相关工具的发展也进展甚微。为应对这一问题，我们提出了几种方法，利用本地新闻网站的内容、一个播送南库尔德语内容的地方广播电台，以及针对拉基语的田野调查。在本文中，我们介绍了这类资源匮乏语言面临的一些挑战，尤其是在书写和标准化方面的挑战；同时也探讨了如何获取数据来源，以及如何对手写内容进行回溯数字化，从而为南库尔德语和拉基语创建语料库。此外，我们还结合库尔德语和扎扎-戈拉尼语的其他变体，研究了语言识别任务。[1]

1 引言

语言和语言学数据在记录和保护濒危语言及资源匮乏语言方面发挥着关键作用。数据对于语言技术中的计算方法不可或缺，也使开发语音识别和机器翻译等工具与应用成为可能，从而支持这些语言的复兴并促进其使用。因此，濒危语言及资源匮乏语言的使用者最终有机会将自己的语言和文化遗产传承给后代。尽管近年来自然语言处理（NLP）取得了令人瞩目的进展，尤其是在低资源设置下处理极少量数据方面（Hedderichet al., 2021年），为濒危及资源较少的语言收集数据仍然是一项艰巨的任务。

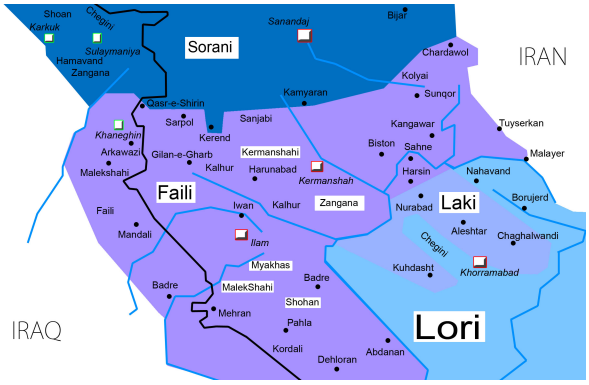


图1：主要使用中库尔德语（深蓝色）、南库尔德语（紫罗兰色）、拉基语（浅紫罗兰色）和卢里语（蓝色）的地区。根据Fattah （2000年）

本文聚焦于南库尔德语（sdh，ISO 639-3），这是库尔德语的主要变体之一，据估计有370万使用者，分布在伊朗的克尔曼沙赫（亦拼作 Kermanshah）省和伊拉姆省，以及伊拉克毗邻的边境地区（Eberhardetal.,2022年）。本文也简要介绍拉基语（lki），这种语言在同一地区有数十万使用者（Aliakbarietal., 2015年）。库尔德语整体上，尤其是南库尔德语和拉基语，长期受到各种歧视性语言政策的影响，造成了有害的社会语言学后果，影响语言态度和传承语维护，例如儿童对南库尔德语掌握不足，以及该语言书面使用有限（Sheyholislami, 2012; Tamlehet al., 2022年）。

因此，南部库尔德语使用者数百年来一直面临波斯语带来的压力：波斯语既是伊朗唯一的官方语言，也是行政上占主导地位的语言，这导致了各种语言转移与变化现象（Sharifi 等人，2013;Weisi, 2021; Yarahmadi, 2022年）。尽管库尔德语的其他变体也未能免受此类政策的影响，但由于其人口更多，且长期以来得到支持，强烈的库尔德民族与政治认同有助于促进语言发展，尤

[1]数据集和模型可从 https://github.com/sinaahmadi/KurdishLID获取

其是对邻近的中库尔德语使用者而言；中库尔德语也被称为索拉尼语（ckb），在伊朗和伊拉克均有使用者（Sheyholislami，2010）。然而，研究表明，南部库尔德语使用者社区同样对库尔德语及其身份认同持积极态度（RezaeiandBahrami，2019; Sheyholislami 和 Sharifi 等人，2016年）。

与其他库尔德语变体，尤其是北库尔德语（kmr，又称库尔曼吉语）和中库尔德语相比，南库尔德语变体和拉基语在语言学或计算语言学领域受到的关注要少得多。此外，南库尔德语和拉基语可用的数字资源相对匮乏，几近于零，而且两者在书写方面都面临实际挑战。另一方面，鉴于该地区的地缘政治条件以及成本和时间上的局限性，研究这些变体的实地情况，即开展语言田野调查，并不总是可行的解决方案。

贡献 本文探讨了如何为南库尔德语和拉基语创建语料库。我们讨论了依托本地内容创作媒体以及田野调查，为代表性不足和濒危语言创建语料库的可行方法。我们的语料库包含超过 200 万个南库尔德语和拉基语词元。除了对语料库进行内在评估外，我们还对语料库进行了定性分析，并将分析扩展到语言识别任务。

2 南库尔德语与拉基语

2.1 语言分类

库尔德人身份认同历经数百年的历史积淀，并深深植根于土地与文化。然而，界定库尔德语的努力一直复杂且充满挑战，并受到政治和社会因素的影响。

尽管很难将库尔德语界定为一种同质语言，而且将其描述为方言连续体还是语言联盟(Jugel 等人，2014年仍有争议，但人们普遍认同，正如Mc-Carus (2009年)、Edmonds (2013年）及其他许多学者所述，北库尔德语、中库尔德语和南库尔德语是库尔德语的三大主要变体。

库尔德斯坦的其他伊朗语支语言，例如扎扎其语（zza）和戈拉尼语（hac）通常被认为不同于库尔德语，尽管两种语言的使用者与邻近的库尔德社群有着密切的文化联系，而且他们也常常认为自己在族群上属于库尔德人（Haig 和 Öpengin，2014年，参见）。尽管如此，这两种语言有时也被称为库尔德语的另外两种方言（Eppler 和 Benedikt，2017年）。此外，将拉基语归类为库尔德语族最南端的变体，是一个有争议的问题。另一方面，学界一致认为，卢里语（也拼作 Lori，lrc/luz）是一种西南伊朗语支语言，尽管人们普遍误以为它是库尔德语的一种变体（Anonby，2004）。不过，卢里语，尤其是拉基语，可能会与邻近的南部库尔德语方言出现趋同现象，反之亦然

本文聚焦于在克尔曼沙赫省使用的南库尔德语变体（本文称之为克尔曼沙希语，也拼作 Kirmaşanî），以及在伊拉克使用的南库尔德语变体。文献将南库尔德语描述为多样化的库尔德语口语集合，可归为若干方言群，其中包括加鲁西方言(2019年，2021年所述的科尔达利方言、卡尔霍里方言和费利语。值得注意的是，本文将“费利语”一词用作对伊拉克边境地区及首都巴格达所使用的若干南部库尔德语方言的统称，同时也承认，将该词用作语言标签存在问题，此处无法进一步讨论。类似地，本文也将所谓的拉基-克尔曼沙赫方言纳入考量；这些变体在Fattah，2000年）中被视为南库尔德语的一部分，但或许更适合描述为混合变体，介于南库尔德语与洛雷斯坦北部、伊拉姆东部的拉基语之间。

2.2 形态句法比较

关于北部库尔德语与中库尔德语变体之间的差异，已有许多研究（Matras，2019; Esmaili 和 Salavati，2013年，参见）。同样，Belelli（2021年，第17页）概述了南库尔德语与其他库尔德语变体之间的主要差异。然而，文献中较少讨论南库尔德语与拉基语之间的差异。

尽管南库尔德语在形态上与北库尔德语和中库尔德语都有相似之处，但它与后者更为接近，因为它没有

词性			北库尔德语	中库尔德语	南库尔德语	Laki
Noun	DEF	M	∅	-eke	-ege, -eke	-e, -ke
		F	∅			
		PL	∅	-ekan	-egan、-ekan、-eğan、-eyle(ge)	-ele
	INDF	M	-ek	-êk	-î、-îg、-îk、-îğ	-ê, -î, -îk
		F	-ek			
		PL	-in	-an , -gel	-eyl、-gel、-ğel、-an	-el
Verb		INF	-in	-in	-in	-in
		PROG	di-	是-，de-	∅，di-，是-	(-是) me-
		SBJV	bi-	bi-	bi-	bi-
形容词		NEG	ne-，na-	ne-，na-，me-	nye-，ne-，na-，nî-	nime-, ne-, nî-
		COMP	-tir	-tir	-tir, -tirek, -tirig	-tir
		SUP	-tirîn	-tirîn	-tirîn	-tirîn

表1：库尔德语和拉基语各变体中词缀的比较。缩写依据莱比锡标注规则（Comrie等人，2008）。为保持一致，所有内容均采用Bedirxan的库尔德化拉丁字母 script。对于缺乏语法性别的变体，名词性词缀合并标注。

性别和格的形态标记。此外，南库尔德语在库尔德语变体中独树一帜，不呈现时态敏感型配列形式，这与北库尔德语和中库尔德语的作格特性不同。另一方面，文献中较少讨论南库尔德语与拉基语之间的差异，尽管有观察指出，拉基语—克尔曼沙希语的各种方言构成一个连续体：从 Sahne 城向哈辛方向移动，拉基语特征逐渐增多；反过来说，沿相反方向移动，南库尔德语特征则逐渐增多（见图1）。哈辛方言在形态和词汇方面与“纯正”拉基语最为相似，而 Sahne 方言则与南库尔德语最为接近。

就拉基语而言，Harsini及其他拉基-克尔曼沙希语方言（如Payrawandi、Sahne’i）所具有的典型拉基语特征包括：存在音位/v/，如 *vitin*[2] ，对比南库尔德语*witin*“说”；存在音位/ö/，如*döm* “尾巴”，对比南库尔德语*dom, dim*及其变体；第二人称复数代词的形式为*homa* ；使用非连续直陈式标记*=a ma-* （但Sahne使用*a-* ，与某些南部库尔德语方言相同）；使用（后）动词小品词，而非库尔德语常见的前置动词成分，例如使用*ör* 代替*hał* “向上”；使用不同的附置词形式，例如*va*“向、在”，对比南库尔德语*wa*“向”、*la/da* “在”；以及使用反身标记*wiž* ，而非南库尔德语的*xwa* 及其变体。

另一方面，Harsini 和其他拉基语——克尔曼沙赫方言有第二人称单数和复数动词词尾的一种形式*-ī(t)/-ītin*，与拉基语的*-ī(n)/-īnān, -īnō(n)* （以及同形的系词黏着形式）不同；第三人称复数黏着代词的一种形式 *=yān* 也不同于拉基语的 *=ān, =ō(n)*。此外，所有拉基语-克尔曼沙希语方言都与南库尔德语一样，在及物动词过去时的变位中没有施事形式；而施事形式原本是拉基语和中库尔德语的显著特征。表格1 总结了一些常见词缀。

2.3 词汇差异

南库尔德语和拉基语中，有一系列词汇是拉基语所特有的，其中包括*āyl* “孩子”，*pit* “鼻子”，*lam* “胃”，*sīr*“吃饱的”，*goǰar* “小的”；与之对应的南库尔德语分别为*minał*，*lūt*、*zik*、*tīr*、*büčik/g*（AliyariBabol-ghani，2019年）。需要指出的是，由于社会语言学和地缘政治条件的影响，南库尔德语和拉基语与几乎所有其他地区语言一样，历史上一直容易从占主导地位的语言中借用词汇，尤其是波斯语和阿拉伯语。

2.4 书写

尽管目前用于书写库尔德语的两种主要文字系统，即以拉丁字母为基础的‘*Hawar*’或‘*Bedirxan*’文字系统，以及中库尔德语使用的波斯-阿拉伯文字，也经过调整用于书写南库尔德语，并采用了 <ڕ> (U+06CA)等不同字形，但这些文字系统在南库尔德语使用者中并不普及，他们在实际书写中依赖行政上占主导地位的——

[2]本节所采用的转写遵循（Belelli，2021年）。

语言的书写系统，即波斯语或阿拉伯语的书写系统（Ahmadi等人， 2019;Fil- ippone等人，2022年）。同样，拉基语也没有标准的文字系统或正字法。

因此，这进一步加剧了这种情况的复杂性：收集到的语料库应采用定制方式书写，或采用一种关系密切的语言的文字系统；在此情形下，即中库尔德语。

3 相关研究

尽管库尔德语是一种资源稀缺语言，但过去几年里，它在语言技术领域日益受到关注，出现了库尔德语处理工具包（Ah-madi，2020b）、Google 翻译等服务[3]，以及自然语言处理领域的 FLORES-101（Goyal etal.，2022）和 NLLB（Costa-jussàetal.，2022）等模型和基准。不过，这些资源仅涵盖北库尔德语和中库尔德语，既不包括南部变体，也不包括拉基语。

同样，维基百科作为记录语言的重要资源，目前仅提供北库尔德语和中库尔德语版本，尚不支持南库尔德语和拉基语，以及其他邻近的资源匮乏语言戈拉尼语和卢里语。Ahmadi等人（2019）研究了库尔德语变体现有的词汇资源，并如图2所示，发现现有的 71 部以电子版和印刷版形式提供、涵盖库尔德语、拉基语和扎扎-戈拉尼语的词典及术语资源中，只有 13.6% 的内容涉及南库尔德语变体或拉基语。

此前，南库尔德语的一些语言学方面已得到研究，例如音系学（KordZafaranlu Kambuziya and Sobati，2014年）、类型学（Dabir-Moghaddam，2012年）、形态学（Belelli，2022年）和方言学（Fattah，2000年）。Belelli（2021年）研究了拉基语，描述了它与南库尔德语之间复杂的关系，并通过田野调查收集词汇表和语料库，记录了一种拉基语变体。

关于语言技术资源，Azin 和 Ahmadi(2021年) 创建了一个基于 Ontolex-Lemon 的电子词典，收录了 14,326 条南库尔德语变体词条，并包含拉基语和卢里语。在这一资源中，词条采用库尔德语常用的两种文字系统表示；尽管拉丁字母拼写法在南库尔德语中并不常用。此外，

[3] https://translate.google.com

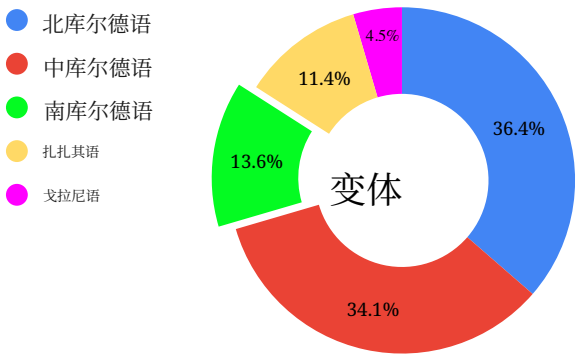


图2：现有库尔德语变体词典资源的占比，其中只有 13.6%（<10 条参考文献）关注南库尔德语和卢里语。

该词典还提供波斯语和中库尔德语译文。据我们所知，这是唯一一部规模可观的南库尔德语电子资源。类似地，Amani 等人(2021年) 收集了用于库尔德语口语方言识别的音频样本，素材来自广播和电视内容，其中南库尔德语音频共收集了 11 小时。表2 概述了库尔德语及其密切相关的语言，以及此前为这些语言创建的一些主要语料库。

除了语言记录之外，语料库还是许多其他应用领域的重要资源，例如语言学习与教学（Tribble，2015年）、机器翻译和句法分析。据我们所知，目前尚未为南库尔德语和拉基语创建规模可观、使用两种规范化库尔德语文字系统之一书写的语料库。

4 方法论

为弥补南库尔德语和拉基语数据不足的问题，我们采用了三种方法，详见本节。

4.1 广播节目

鉴于当地媒体的存在，我们选择了伊朗克尔曼沙赫省的一家地方广播电台，该省的大多数居民以南库尔德语为母语。应我们的请求，我们收集到了一组以南库尔德语克尔曼沙希语变体写成的广播节目手稿。这些脚本涵盖教育、文化和日常话题，主要面向农村地区的听众。因此，其中使用了丰富的南库尔德语本土词汇，几乎没有（如果有的话）语码转换成波斯语或其他语言的情况。

语言	639-3	维基百科	常用文字	语料库
北库尔德语（库尔曼吉语）	kmr	ku	拉丁字母，中库尔德语	(Esmaili 和 Salavati，2013;Ataman, 2018;Matras, 2019年)
中库尔德语（索拉尼语）	ckb	ckb	中库尔德语，拉丁字母	(Esmaili等人，2013;Abdulrahman等人，2019; Veisi等人，2020; Ahmadi等人，2020; Matras，2019年)
南库尔德语	sdh	-	中库尔德语，波斯语	(Fattah, 2000年)
戈拉尼语	hac	-	中库尔德语	(Ahmadi, 2020年a)
扎扎其语	zza	diq	拉丁字母	(Ahmadi，2020a)

表2：按照流行程度排列的库尔德语变体、扎扎其语和戈拉尼语的描述，以及部分现有语料库和文字系统。中库尔德语文字指中库尔德语中通常使用的库尔德化波斯-阿拉伯字母。

大量从波斯语借用词汇。

广播节目数据集由18个为克尔曼沙赫市一家地方广播电台撰写的脚本组成。这些脚本用于谈话节目和短喜剧，并由同一广播频道播出。脚本是为时长10至15分钟的节目撰写的。大多数节目采用对话形式，因此该数据集非常适合未来的话语分析研究。

原始场景由人工编写，使用波斯文书写系统与正字法。我们请三位南部库尔德语使用者使用中库尔德语波斯-阿拉伯文字输入这些场景。这使我们能够将这些数据与用其他库尔德语变体撰写的材料进行比较。随后，我们对手动输入的数据进行审查，以检查输入者所用书写形式是否存在不一致之处。

4.2 新闻文章

在我们的第二种方法中，我们采用Ahmadi(2020a的方法，从新闻网站抓取内容，以记录伊拉克境内使用的南库尔德语变体。我们找到了一家本地新闻网站[4]，该网站以包括费利语在内的几种语言发布新闻文章。我们总共抓取了15,985篇文章，以HTML格式保存并转换为文本。之后，我们进行文本预处理，使用正则表达式统一字符编码，清理原始文本，删除电子邮件地址等私人信息及文本格式，并根据文章主题对原始文本进行分类，主要分为文化、政治和库尔德斯坦类别。

作为元数据，我们提供所收集文章的来源、主题、标题和日期。

[4]https://shafaq.com

4.3 田野调查

最后，我们通过田野调查记录拉基语，并在伊朗西部克尔曼沙赫省哈辛市使用的语言变体中创建口语文本语料库。该变体属于所谓的克尔曼沙希语（或克尔马沙尼拉克语）方言群，被认为介于南库尔德语和洛雷斯坦拉克语之间（Belelli，2021年）。哈辛文本语料库的内容在类型学上具有一致性，包含七篇传统叙事——五篇民间故事和两则轶事——均为独白形式，录自四位母语者（三位女性和一位男性），他们来自哈辛或附近的 Parive 村。文本依照基于伊朗语言学传统的常规转写系统进行人工转写，划分为带编号的标注单元，并翻译成英语。其中一篇文本还进一步做了逐词素注释。

由于拉基语没有标准书写系统或正字法，使用波斯文书写系统或库尔德化的拉基语文字系统仍属于可选方案，而非惯常选择。尽管如此，鉴于语音转写具有一致性，对语料库进行转写仍是可行的。

5 结果与分析

在本节中，我们对语料库开展内在评估，并 प्रस्तुत定性分析。同时，我们还将分析扩展至语言识别任务。

5.1 定量分析

收集到的数据包含 16,003 份以南库尔德语各变体写成的文档，以及七篇拉基-克尔曼沙希语叙事文本。表3 列出了

我们收集的语料库中的文章数量、词元数、类型数和类型字符数。为了计算类型（即唯一词元），我们排除了标点符号、数字，以及暂时被标记为语码转换的句子，例如阿拉伯语宗教引文或波斯语诗歌。此外，我们使用正则表达式进行词元化。

由于所选变体的词汇有很多共同之处，我们还计算了平均词类型长度，作为词形态复杂度的指标。尽管克尔曼沙希语和拉基语数据规模较小，可能不足以揭示这些变体形态结构的复杂性，但克尔曼沙希语的平均词长为 6.57，拉基语为 6.45，这似乎与费利语词类型的平均长度 8.8 相差甚远。与库尔德语和扎扎-戈拉尼语的其他变体相比，南库尔德语的词形总体上似乎更长。根据 (2020年a)，北库尔德语和中库尔德语的平均长度分别为 4.8 和 5.6。同样，扎扎其语和戈拉尼语的平均长度分别为 4.84 和 5.50。

我们认为，词长的这一显著差异可能源于：a) 南库尔德语语料库的正字法，尤其是费利语文本；b) 多词表达的书写惯例，例如将روشونکریاگ بیسه (*bîserûşûnkiryag*)（“注定要遭殃的”）写作ناگ بی-سەر-و-شوون-کرێ；以及 c) 过度串接词语，例如将گورانیچرلوبنانی 写作，而不是 گورانیچر لوبنانی (*goranîçirrLubnanî*)（“黎巴嫩歌手”）。我们还注意到，连词 و (*û*)（“和”）以及诸如 له(*le*)（“在”）这样的介词，有时会与前一个或后一个词合并，例如将 پەلامارەدەرێلەو 写作，而不是 پەلامارەدەرێلەو (*pelamardereyle û*)（“攻击者和”）；又如将 لەشەقامێک 写作，而不是 لەشەقامێک (*le şeqamêk*)（“在一条街上”）。更重要的是，如表 1所示，南库尔德语的词缀比北库尔德语和中库尔德语的词缀更长，因此平均词长更高。

此外，我们还计算了Pewan语料库中北库尔德语和中库尔德语的秩-规模分布（Esmaili等人，2013年）以及扎扎-戈拉尼语语料库（Ahmadi，2020a）的秩-规模分布，并将其与我们的南库尔德语数据（合并克尔曼沙希语和费利语）一同进行分析。根据齐夫定律（齐夫，1999年），在这种分布中，“词语长度往往与其相对频率成反比”。图3 展示了这一现象：每个语料库的曲线都从最高频词开始（以点表示

数量（#）	克尔曼沙希语	费利语	Laki
文章	18	15,985	7
词元	10,127	2,182M	6,340
类型	3,248	179,208	2,074
字符	21,359	1,591M	13,378
平均长度	6.57	8.8	6.45

表3：南库尔德语各变体的收集数据统计（M 指百万）。字符数和平均长度根据类型计算。

），随后，秩为10至10000的词语词频平稳下降，最后，大多数词语以较低词频出现在底部区段（<10）。由于这种分布需要规模相对较大的语料库才能成立，我们无法纳入拉基语数据。

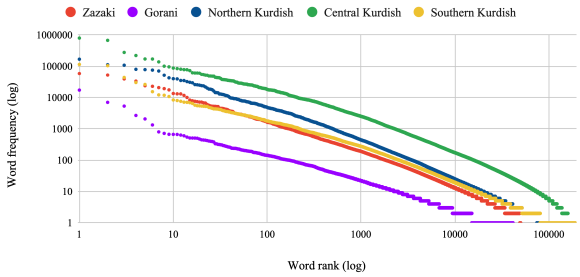


图3：北库尔德语和中库尔德语的 Pewan语料库、一个扎扎-戈拉尼语语料库，以及我们的南库尔德语语料库（合并了克尔曼沙希语和费利语）的齐夫分布。

5.2 定性分析

从定性的角度来看，由于收集的数据属于不同的文本体裁，人们显然会预期文本结构、词汇选择和正式程度存在差异。然而，就伊拉克和伊朗境内使用的南库尔德语变体而言，最显著的差异与可归类为从其他区域语言借入的词语的使用有关。有趣的是，伊拉克使用的南库尔德语变体也倾向于采用中库尔德语词汇，尤其是在术语方面；而伊朗使用的变体则更多采用波斯语词汇。

根据为费利语收集的新闻文章的文本结构，每篇文章都有一个标题，以主动语态简明概括内容。标题集为语篇分析和语料库的进一步语言学研究提供了极佳的资源。

另一方面，为克尔曼沙希语收集的数据包含为广播节目撰写的独白和对话。喜剧节目中的对话采用双向交谈的形式，由配音演员围绕社会和文化议题展开。节目中使用的非正式语言模拟了两位克尔曼沙希语使用者之间的真实互动，为今后的会话分析提供了机会。同样，拉基语数据中的叙事内容也提供了分析民间故事的有用信息。

此外，齐夫定律还指出，由于经济因素，一种语言中出现频率最高的词往往较短（Sigurd等人，2004）。表4 列出了所选语料库中出现频率最高的词。尽管许多高频词少于三个字符，例如*û* “和”，*li/le/ce/de*“在”，以及*bo/ara/ařā*“为了”，但许多其他词，如库尔德斯坦和伊拉克也频繁出现，这反映了文本所涉及的主题，同时也表明了语料库存在偏向且领域多样性不足。这也受到语言正字法的影响：扎扎其语中的后置词*ra* 和北库尔德语中的*de* 频繁出现，而中库尔德语中的对应形式，如*da* 和*ewe* 则较少出现。尽管如此，拉基语数据中出现频率最高的词体现了叙事文本中的一些元素，例如*muše* “IND-SAY.PRS-3SG”。

尽管南库尔德语各变体与拉基-克尔曼沙希语共享许多语言特征，但两者之间的差异不容忽视。正如前文所述，语言数据稀缺，且这一库尔德语分支缺乏书写系统，这些都是我们至今仍无法清晰了解其现存次级变体在书面形式上的差异的原因。

5.3 语言识别

语言识别是指检测句子所使用语言的任务。这项任务广泛应用于自然语言处理的许多下游应用，例如情感分析（Vilares等人，2017年）、文本摘要（Kanapala等人，2019年）、多语言文档和万维网上的混合语检测（Bhargava等人，2016年）以及机器翻译（Sefara等人，2021年）。尽管此前已有研究针对部分库尔德语变体（如中库尔德语）开展语言识别（Malmasi，2016年），但尚未有研究涵盖所有库尔德语变体。

除了从我们的语料库中提取的句子外我们还从现有语料库中收集了其他库尔德语变体的3000个句子，具体如下：来自Pewan语料库的使用波斯-阿拉伯文字书写的中库尔德语和使用拉丁字母书写的北库尔德语（Esmaili等人，2013年），来自Wergor语料库的使用拉丁字母书写的中库尔德语（Ahmadi，2019年），以及来自Ahmadi (2020a）的扎扎其语和戈拉尼语句子。鉴于北库尔德语的某些变体也使用波斯-阿拉伯文字书写，尤其是在伊拉克库尔德斯坦，我们还从以波斯-阿拉伯文字书写北库尔德语的在线论坛和网站收集句子。此外，我们注意到，扎扎其语专属维基百科页面所用的文字系统和正字法[5] 与Ahmadi (2020a）语料库中使用的不同; 后者完全采用北库尔德语中通行的“*Hawar*”或“*Bedirxan*”体系（Littell等人，2016年），而前者则受到土耳其拉丁字母的影响。由于拉基语的书写系统在实践中尚未确立，我们未将拉基语数据纳入本任务。为了进一步丰富本任务，我们还纳入了来自Tatoeba数据集的阿拉伯语、波斯语和土耳其语句子。[6]

作为基线系统，我们评估 fastText 的预训练语言识别器（Bo-janowskietal.，2017年），该识别器可识别176种语言，包括北库尔德语、中库尔德语和扎扎其语，分别使用 `kmr`、 `ckb` 和`diq` 识别器。此外，我们训练了分类器，其目标类别（即语言标签）包含文字系统代码，例如，使用 `ckbarab` 和`ckblatn` 来区分分别以波斯-阿拉伯文字和拉丁字母书写的中库尔德语（`ckb`）文本。类似地，我们还考虑了仅依据语言代码汇总标签的分类场景。为此，我们使用 fastText 训练模型，超参数如下：25个训练轮次，词向量维度为64，字符*n*n元模型的最小和最大长度分别为2和6，学习率为1.0，并采用分层 softmax 作为损失函数。

表5 展示了我们针对所选变体和文字系统进行语言识别的实验结果。尽管预训练fastText模型表现不佳，`lid.176` 这主要是因为它未在我们的目标语言上接受过训练。结果表明，我们训练的模型

[5] https://diq.wikipedia.org
[6] https://tatoeba.org

北库尔德语	中库尔德语	南库尔德语		Laki	戈拉尼语	扎扎其语
		费利语	克尔曼沙希语			
和 （和）	从；在 （从，在）	是 （是）	和 （和）	说 （直陈式-说-现在时-第三人称单数）	和 （和）	在 （在）
ku （那个）	和 （和）	和 （和）	和 （和）	这 （这个、这些）	在 （在.....里面）	和 （和）
从；在（从、在）	为了 （为了）	那 （那）	*le* （在）	*ye* （一个）	*be* （向；与）	*ke* （即）
bi （用；向）	*be* （与；向）	和 （和）	村庄 （村庄）	*arā* （为了）	*ke* （那）	*ra*
di （在）	*ke* （那）	*era* （为了）	*naw* （在;name中）	*va* （到）	*pey* （为了）	*bi* （用）
ji （从）	那个 （那个）	那个 （那个）	和 （和）	去 （直陈式-去.现在时-第三人称单数）	*y*	和 （我们）
de	库尔德斯坦	*kird* (使.....做了)	*ta* (直到)	*ya*(这个；这个人)	*ta* (直到)	*xo* (自己)
jî (也)	*Iraq*	wit（IND-SAY.PST-3SG）	*ê* （这个）	*a* （是; 那个）	*i* （这个）	*zî* （也）
库尔德斯坦	*em* （这个）	*herêm* （地区，所属地区）	*i* （这个）	*make* （直陈式非过去时-做-第三人称单数）	库尔德斯坦	*yê*
伊拉克	地区 （.....的地区）	库尔德斯坦	之后 （之后）	我（我，我的）	每 （每个）	我的 （我的）
herêma （地区的）	*serokî* （总统的）	*ta* （直到）	*bün* （IND-BE.PST-3PL）	*nām* （名称）	*Turkyay* (土耳其)	*o* （那，它）

表 4：北库尔德语、中库尔德语和扎扎-戈拉尼语语料库中出现频率最高的 10 个词，以及我们收集的南库尔德语和拉基-克尔曼沙希语数据。除了介词和连词等常见的功能词外，许多词都与文本的主题相关，例如库尔德斯坦 和伊拉克。

测量	lid.176	我们的模型		
		语言代码	语言与文字代码	SDH-非常规
精确率	0.0552	0.969	0.9638	0.25
召回率	0.0674	0.971	0.9636	0.126
F_1	0.06	**0.97**	0.9634	0.168

表 5：在分类标签中包含和不包含文字代码（arab,latn）时的语言识别结果。非常规指识别以波斯文书写系统书写的南库尔德语文本，而非库尔德语。我们的模型优于基线（预训练fastText）。各项指标采用算术平均值计算（也称宏平均值或非加权平均值）。

在两种设置下都表现良好：一种是仅为分类任务提供语言代码，如ckb ；另一种是提供文字代码，如 ckbarab。我们还在南库尔德语数据上评估了模型。该数据采用波斯文书写系统与正字法书写，在与中库尔德语波斯-阿拉伯文字系统协调统一之前便是如此。阴性$_1$ 度量值为0.168，反映了在这种噪声环境下执行此任务的难度。表A.1 和图A.1中给出了一些预测示例及其热图。

6 结论

总体而言，数据，尤其是语料库，为在语言技术领域保存和推广濒危及资源匮乏语言奠定了基础。本文讨论了三种低资源和资源匮乏语言的数据收集与语料库创建方法，涉及在伊朗克尔曼沙赫省使用的南库尔德语变体，以及在伊拉克使用的费利语变体。克尔曼沙希语数据集来自当地广播电台，并通过抓取新闻网站收集；而拉基语的数据则通过田野调查收集。尽管面临诸多挑战，田野调查似乎是为一种几乎没有网络存在的语言收集数据的唯一解决方案。我们提出的方法也可供其他数据有限且无法开展田野调查的资源匮乏语言采用。最后，我们从定量和定性角度进行了简要分析，并评估了使用不同文字系统的库尔德语和扎扎-戈拉尼语的语言识别效果。我们的模型有助于检测文本并收集更多数据。

关于未来工作，可以探索数据驱动方法，以阐明所选语言和变体之间的各种语言差异。此外，由于我们的目标语言一直面临语言同化的威胁（Hasanpoor，1999年），尤其是受到从波斯语、阿拉伯语和土耳其语中进行词汇借用的影响，因此出现了一个新的问题，即如何确定词汇借用。我们认为，借助我们的数据，未来可以进一步研究词汇借用检测（Miller et al.，2020年）。所收集的语料库可以为语言技术和跨语言研究的进一步发展铺平道路。最后，未来还可以为这些语料库添加其他任务的标注，尤其是词性标注和命名实体识别。

7 局限性

当前研究的主要局限性之一，是所收集数据的规模较小，其中克尔曼沙赫语和拉基语的词元数量均少于20,000个。因此，有必要扩展当前数据，以便能够基于语料库以有意义的方式分析这些语言（Davies，2018年）。还可以扩展定性分析，根据文本类型以及变体和语言，考察句子和词长偏好。

为统一拉基语数据并使其与其余南库尔德语语料库具有可比性，需要对拉基语语料库进行转写。然而，要采用一种作为通行规范的书写系统，还需要相关语言社群进行更多讨论。

致谢

Sina Ahmadi 和 Antonios Anastasopoulos 获得美国国家科学基金会通过 DEL/DLI 项目资助 BCS-2109578 提供的慷慨资助。作者也感谢匿名评审者。

参考文献

Roshna Abdulrahman、Hossein Hassani 和 Sina Ahmadi。2019年。为一种资源稀缺语言构建细粒度语料库：库尔德语的案例。载于*2019*年自然语言处理拓展研讨会论文集，第106–109页。

Sina Ahmadi。2019年。基于规则的库尔德语文本转写系统。*ACM* 亚洲低资源语言信息处理汇刊，18(2):18:1–18:8。

Sina Ahmadi。2020a。构建扎扎-戈拉尼语族语料库。载于第七届相近语言、变体和方言的自然语言处理研讨会论文集，*VarDial@COLING 2020*，西班牙巴塞罗那（线上），*2020*年*12*月*13*日，第70–78页。国际计算语言学委员会（ICCL）。

Sina Ahmadi。2020b。KLPT——库尔德语处理工具包。载于第二届自然语言处理开源软件研讨会（*NLP-OSS*）论文集，第72–84页，线上。计算语言学协会。

Sina Ahmadi、Hossein Hassani 和 DabanQ. Jaff。2020年。利用多语种新闻网站构建库尔德语平行语料库。*CoRR*，abs/2010.01554。

Sina Ahmadi、Hossein Hassani 和 John P McCrae。2019年。《迈向库尔德语电子词典编纂》。载于第六届双年度电子词典编纂会议（*eLex*）。eLex 2019。

Mohammad Aliakbari、Mojtaba Gheitasi 和 Erik Anonby。2015年。《伊朗伊拉姆省的语言分布》。《伊朗研究》，48(6):835–850。Salman Aliyari Babolghani。2019年。《拉克语是一种库尔德语方言吗？》。《纪念Adriano V. Rossi的伊朗研究》，第3–20页。Arash Amani、Mohammad Mohammadamini 和 Hadi Veisi。2021年。《使用X-Vector 说话人嵌入识别库尔德语口语方言》。载于*Speech and Computer*：第*23*届国际会议，*SPECOM 2021*，俄罗斯圣彼得堡，*2021*年*9*月*27–30*日，论文集*23*，第50–57页。Springer。

Erik Anonby. 2004. 库尔德语还是卢里语：伊朗洛雷斯坦省拉基语的身份认同之争。*Kurdische Studien*，4(5):7–22。

Duygu Ataman. 2018年. Bianet：土耳其语、库尔德语和英语平行新闻语料库。*arXiv* 预印本 *arXiv:1805.05095*。

Zahra Azin 和 Sina Ahmadi. 2021年. 为资源匮乏的库尔德语南部变体创建电子词汇库。第七届双年电子词典学会议（*eLex 2021*）论文集。

Sara Belelli. 2019年. 构建南库尔德语方言学：从何处入手。库尔德语言学当前议题，1:73。

Sara Belelli. 2021年. 哈辛拉基语变体：语法、文本、词汇。班贝格大学出版社，班贝格。

Sara Belelli. 2022年. 库尔德语过去时范畴的跨方言论述，特别以南库尔德语为例。载于库尔德语方言的结构与类型变异，第239–290页。Springer。

Rupal Bhargava、Yashvardhan Sharma 和 Shubham Sharma。2016年。混合文字印度语句的情感分析。载于*2016*年计算、通信与信息学进展国际会议，*ICACCI 2016*，印度斋浦尔，*2016*年*9*月*21*日至*24*日，第524–529页。IEEE。

Piotr Bojanowski、Edouard Grave、Armand Joulin 和 Tomas Mikolov。2017年。利用子词信息丰富词向量。计算语言学协会汇刊，5:135–146。

Bernard Comrie、Martin Haspelmath 和 Balthasar Bickel。2008。《莱比锡标注规则：逐行逐语素注释惯例》。马克斯·普朗克进化人类学研究所语言学系 & 莱比锡大学语言学系。检索于一月，28：2010。

Marta R. Costa-jussà、James Cross、Onur Çelebi、Maha Elbayad、Kenneth Heafield、Kevin Heffernan、Elahe Kalbassi、Janice Lam、Daniel Licht、Jean Maillard、Anna Sun、Skyler Wang、Guillaume Wenzek、Al Youngblood、Bapi Akula、Loïc Barrault、Gabriel Mejia Gonzalez、Prangthip Hansanti、John Hoffman、Semarley Jarrett、Kaushik Ram Sadagopan、Dirk Rowe、Shannon Spruit、Chau Tran、Pierre Andrews、Necip Fazil Ayan、Shruti Bhosale、Sergey Edunov、Angela Fan、Cynthia Gao、VedanujGoswami、Francisco Guzmán、Philipp Koehn、Alexandre Mourachko、Christophe Ropers、Safiyyah Saleem、Holger Schwenk 和 Jeff Wang。2022年。没有语言被落下：扩大以人为中心的机器翻译规模。*CoRR*，abs/2207.04672。

Mohammad Dabir-Moghaddam。2012年。《语言类型学：伊朗视角》。《世界语言杂志》，13(1):31–70。

Mark Davies。2018年。《词汇和语义变异的基于语料库的研究：语料库规模和语料库设计同等重要》。载于《从数据到证据：英语研究》，第66–87页。布里尔出版社。

David M Eberhard、Gary F Simons 和 Charles D Fen- nig。2022年。《民族语》：世界语言。达拉斯：国际语言学暑期学院。

Alexander Johannes Edmonds。2013年。《库尔德语方言》。海德堡鲁普雷希特-卡尔斯大学。

Eva Eppler 和 Josef Benedikt。2017年。采用感知方言学方法研究库尔德语变体的语言变异与空间分析。《语言地理学杂志》，5(2)：109–130。

Kyumars Sheykh Esmaili、Donya Eliassi、Shahin Salavati、Purya Aliabadi、Asrin Mohammadi、So- mayeh Yosefi 和 S hownem Hakimi。2013年。构建索拉尼库尔德语测试语料集。载于*2013*年*ACS*国际计算机系统与应用会议（*AICCSA*），第1–7页。IEEE。

Kyumars Sheykh Esmaili 和 Shahin Salavati。2013年。索拉尼库尔德语与库尔曼吉库尔德语：实证比较。载于计算语言学协会第*51*届年会论文集（第*2*卷：短文），第2卷，第300–305页。

Ismail Kamandar Fattah。2000年。《南库尔德语方言：语言学与方言学研究》（《伊朗学报》第*37*卷）。皮特斯出版社，鲁汶。

Ela Filippone、Sara Belelli 和 Matteo De Chiara。2022年。《被“笔”与“剑”分隔：从贝迪尔汗兄弟和*Taufiq Wahby*到南库尔德语的库尔德语正字法编纂》（被“笔”与“剑”分隔：从贝迪尔汗兄弟和Taufiq Wahby到南库尔德语的库尔德语正字法编纂）。第93–106页。

Naman Goyal、Cynthia Gao、Vishrav Chaudhary、Peng- Jen Chen、Guillaume Wenzek、Da Ju、Sanjana Kr- ishnan、Marc'Aurelio Ranzato、Francisco Guzmán 和 Angela Fan。2022年。面向低资源语言和多语种机器翻译的Flores-101评测基准。《计算语言学协会汇刊》，10：522–538。

Geoffrey Haig 和 Ergin Öpengin。2014年。《库尔德语特刊导言——研究综述与批判》。库尔德研究，2(2):99–122。

Jafar Hasanpoor。1999年。《标准索拉尼语中欧洲语、波斯语和阿拉伯语借词研究》。乌普萨拉大学。

Michael A. Hedderich、Lukas Lange、Heike Adel、Jannik Strötgen 和 Dietrich Klakow。2021年。低资源场景下自然语言处理近期方法综述。载于《*2021*年计算语言学协会北美分会会议论文集：人类语言技术》，第2545–2568页，在线出版。计算语言学协会。

Thomas Jugel 等人。2014年。论库尔德语语言史。库尔德研究，2(2)：123–142。

Ambedkar Kanapala、Sukomal Pal 和 Rajendra Pamula。2019年。法律文书文本摘要：综述。人工智能评论，51(3)：371–402。

Aliyeh Kord Zafaranlu Kambuziya 和 Elham Sobati。2014年。卡尔霍里库尔德语方言中的辅音音系过程。语言相关研究，5(1)：191–222。

Patrick Littell、David R Mortensen、Kartik Goyal、Chris Dyer 和 Lori Levin。2016年。西伊朗语支中的桥接语言大小写推断：索拉尼语、库尔曼吉语、扎扎其语和塔吉克语。载于第十届语言资源与评测国际会议论文集（*LREC'16*），第3318–3324页。

Shervin Malmasi。2016年。索拉尼库尔德语的次方言差异。载于第三届相近语言、变体和方言自然语言处理研讨会论文集（*VarDial3*），第89–96页。

Yaron Matras。2019年。重新审视库尔德语方言地理：曼彻斯特数据库的发现。库尔德语言学当前议题，1：225。

Ernest N. McCarus。2009年。库尔德语。*Windfuhr, Gernot*（编）。伊朗语族。伦敦：*Routledge*，1：587–633。

John E Miller、Tiago Tresoldi、Roberto Zariquiey、César A Beltrán Castañón、Natalia Morozova 和 Johann-Mattis List。2020年。使用词汇语言模型检测单语词表中的借词。*PLOS ONE*，15(12)：e0242709。

Saeed Rezaei 和 Ava Bahrami。2019年。伊朗伊拉姆市对库尔德语的态度。伊朗境内外的伊朗语言社会语言学：关于波斯语、阿塞拜疆语和库尔德语的

案例，第77–106页。

Tshephisho J. Sefara、Skhumbuzo G. Zwane、Nelisiwe Gama、Hlawulani Sibisi、Phillemon N. Senoamadi 和 Vukosi Marivate。2021年。面向嵌入语言识别的低资源语言 Transformer 机器翻译。载于信息通信技术与社会会议，*ICTAS 2021*，线上活动／南非德班，*2021*年*3*月*10*日至*11*日，第127–132页。IEEE。

Shahla Sharifi、Mahmoud Elyasi 和 Amir Karimi Pour。2013年。60名库尔德语—波斯语双语者所产库尔德语叙事中的语言变化分析。《亚洲社会科学与人文科学杂志》，2。

Jaffer Sheyholislami。2010年。身份认同、语言与新媒体：库尔德语格。《语言政策》，9(4)。

Jaffer Sheyholislami。2012年。伊朗的库尔德语：受限制和受控的宽容案例。国际语言社会学杂志，2012(217):19–47。

Jaffer Sheyholislami 和 Amir Sharifi。2016年。“最难维持的”：库尔德语作为美国的传承语。《国际语言社会学杂志》，2016(237)：75–98。

BengtSigurd、MatsEeg-Olofsson 和 JoostVan Wei- jer。2004。词长、句子长度和词频——重新审视齐夫定律。《语言学研究》，58(1)：37–52。

Hadis Tamleh、Saeed Rezaei 和 Nettie Boivin。2022年。伊朗克尔曼沙赫的库尔德语—波斯语家庭语言政策。《多语》，41(6)：743–767。

Christopher Tribble。2015年。教学与语言语料库。语言语料库对数据驱动学习的多重助益，69：37–62。

Hadi Veisi、Mohammad MohammadAmini 和 Hawre Hosseini。2020 年。迈向库尔德语处理：AsoSoft 文本语料库的收集与处理实验。《人文数字学》，35(1)：176–193。

David Vilares、Miguel A. Alonso 和 Carlos Gómez- Rodríguez。2017年。多语言环境中的监督式情感分析。《信息处理与管理》，53(3):595–607。

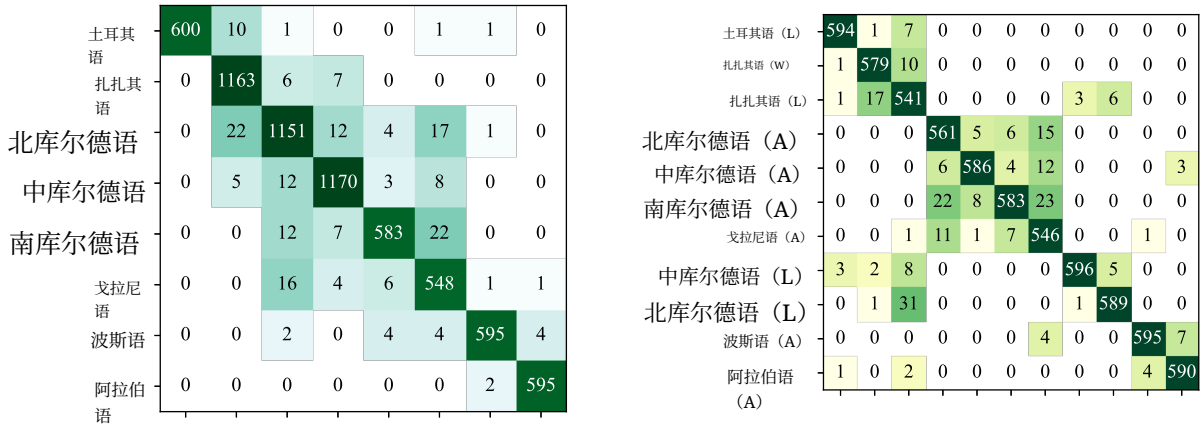
Hiwa Weisi。2021年。伊朗多语言环境中卡尔胡里库尔德语使用者的语言优势与语言转移：语言自杀还是语言灭绝？《语言问题与语言规划》，45(1):56–79。

贾瓦德·亚拉赫马迪。2022年。伊朗卡尔胡里库尔德语使用者的语言变化：再次尝试还是徒劳？语用学与社会，13(2):322–340。

G.K. Zipf。1999年。《语言的心理生物学：动态语文学导论》。认知心理学]。Routledge。

语言	文字系统	预测		句子
		lid.176	我们的	
北库尔德语	Bedirxan	KU	KU	Evî jêderî 说，Reşave 村庄和 Kukerê 遭到了此次炮击。
北库尔德语	中库尔德语	CKB	KU	近6个月来，政治力量在选举中取得的成功仍在持续。
中库尔德语	中库尔德语	CKB	CKB	此外，他们报告说，所有医院都收治了伤者。
中库尔德语	Bedirxan	KU	CKB	土耳其议会投票支持欧洲联盟一揽子方案。
南库尔德语（费利语）	中库尔德语	CKB	SDH	今天（星期一），外币和伊拉克货币在本地市场的汇率下跌
南库尔德语（克尔曼沙希语）	中库尔德语	CKB	SDH	愿我们写下的诗句韵律优美，令人欣喜。
南库尔德语（克尔曼沙希语）	波斯语	FA	FA	我们希望在亲爱的克尔曼沙赫省的每个地方，都能听到我们的声音，并感到愉快。
扎扎其语	Bedirxan	KU	ZZA	您正在为德尔西姆人民的权利归还以及政治政策的终结而努力
扎扎其语	维基百科	DIQ	ZZA	阿加里亚基语是一种印度斯坦语。属于孟达语族。
戈拉尼语	中库尔德语	CKB	HAC	拥有较高社会地位和身份的人，例如自由、独立且出众的人。

表 A.1：所选语言中的一些示例，以及 fastText 预训练模型（`lid.176`）的预测结果，并与我们使用 fastText 在所收集数据上训练的模型的预测结果进行比较。这里使用了不同文字系统的北库尔德语（KU）、中库尔德语（CKB）和南库尔德语（SDH），以及戈拉尼语（HAC）和扎扎其语（ZZA）。DIQ 指的是维基百科上用于扎扎其语的文字系统。



（a）使用语言代码进行分类　　（b）使用语言和文字代码进行分类

图 A.1：将文字系统作为标签（右侧）和不使用文字系统（左侧）时，库尔德语变体和扎扎-戈拉尼语的语言识别结果。文字代码以 L、A 和 W 表示，分别对应拉丁字母、阿拉伯语和扎扎其语维基百科。图中标注了分类结果的数量。横轴和纵轴分别表示参考标签和模型预测。