文章编号：1003-0077(2018)02-0081-06

THUUyMorph：维吾尔语形态切分语料库

哈里旦木·阿布都克里木[1]，孙茂松[1]，刘 洋[1]，阿布都克力木·阿布力孜[2]

（1. 清华大学 计算机科学与技术系 智能技术与系统国家重点实验室，清华信息科学与技术国家实验室(筹)，北京 100084）
（2. 清华大学 人文学院 计算语言学实验室，北京 100084）

摘 要：THUUyMorph(Tsinghua University Uyghur Morphology Segmentation Corpus)是由清华大学自然语言处理与社会人文计算实验室构建的维吾尔语形态切分语料库。原始语料从 2016 年的天山网维文版①下载，题材内容包含新闻、法律、财经、生活等。语料库构建步骤为：爬虫、校对原始语料、分句、校对分句、人工和自动形态切分结合、人工标注语音和谐变化现象、人工校对形态切分和语音和谐变化现象。语料库包含 10 596 个文档、69 200 个句子，词语类型为 89 923 个，分为词级和句子级两类标注，开源网址为 http://thuuymorph.thunlp.org/。该研究不仅对维吾尔语语料库的建设具有参考意义，而且为维吾尔语自然语言处理的研究提供了有益的资源。

关键词：THUUyMorph；维吾尔语；形态切分

中图分类号：TP391 **文献标识码**：A

THUUyMorph：An Uyghur Morpheme Segmentation Corpus

Halidanmu Abudukelimu[1], SUN Maosong[1], LIU Yang[1], Abudukelimu Abulizi[2]

(1. State Key Laboratory of Intelligent Technology and Systems, Tsinghua National Laboratory for Information Science and Technology, Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China;
2. Laboratory of Computational Linguistics, School of Humanities, Tsinghua University, Beijing 100084, China)

Abstract: THUUyMorph (Tsinghua University Uyghur Morphology Segmentation Corpus) is an Uyghur corpus with morpheme segmentation annotations. The original corpus is downloaded from Tianshan website in 2016, including news, law, life, etc. Corpus are processed by proofreading of the original corpus, clauses segmentaion and proofreading, manual and automatic annotation for morpheme segmentation, manual annotation of phonetic harmony phenomenon, manual correction of morpheme segmentation and phonetic harmony. The corpus contains 10,596 documents, 69,200 sentences and 89,923 word types, which are annotated at both word-level and sentence-level. The corpus is available at http://thuuymorph.thunlp.org/.

Key words: THUUyMorph; Uyghur; morpheme segmentation

0 引言

形态切分(morphological segmentation)是将一个词切分成形态或语素的结构化预测任务，其输出结果能够帮助提高各种不同应用任务的性能，如自动语音识别[1]、词汇表示学习[2]、机器翻译[3]和句法分析[4-5]等。形态丰富的语言存在大量形态不同的词，造成在执行自然语言处理时出现严重的数据稀疏问题。例如，维吾尔语通过屈折和派生可以生成无限数量的词。Hamkamer[6]认为黏着语建立词典是不可能的。因此，词应该被切分成最小语义单

收稿日期：2017-07-16 **定稿日期**：2017-12-26

基金项目：国家自然科学基金(61331013)；国家“863”高技术项目(2015AA015407)

① http://uy.ts.cn/

元——语素。例如，维吾尔语词باغچىلاشتۇرماقچىمىز（Latin：baghchilashturmaqchimiz，释义：我们准备建园林），切分成语素是باغچىلاش#تۇر#ماقچى#مىز[7]。因此，形态切分是维吾尔语自然语言处理领域基础且重要的任务。

深度学习在自然语言处理领域中广泛应用，形态切分工作也取得了极大进展，实现了从规则和传统统计方法向神经网络方法的跨越[8]。然而，当前的深度学习技术主要是有监督的学习，深度学习的成功运用前提是先具有一定规模的标注语料[9]。

维吾尔语在语料库建设方面已做了大量的工作。新疆大学吐尔根·依布拉音等[10-12]和新疆师范大学的玉素甫·艾白都拉等[13-14]都已构建了百万词次的维吾尔语词法分析语料库，并分别在这些语料库基础上进行了词法、句法及面向具体任务的标注等。除此之外，文献[15]构建了FrameNet，文献[16]建立了语法信息词典，文献[17]建立了小规模命名实体关系语料库。虽然当前已有了相当规模的维吾尔语语料库，但是还没有可公开使用的维吾尔语形态切分语料库。

本文建立的形态切分语料库——THUUyMorph，分为词级和句子级两种，可用于维吾尔语有监督、半监督、无监督的形态切分，以及维吾尔语分词、词干提取等任务。在建立过程中本文参考了Ryan Cotterell的工作[18]。建立和公开的维吾尔语形态切分语料库的开源网址为：http://thuuymorph.thunlp.org/。该研究不仅对维吾尔语语料库的建设具有参考意义，而且为维吾尔语自然语言处理研究提供了有益的资源。

1 研究背景

1.1 维吾尔语形态切分的特点

世界上语言分类包括：孤立语、屈折语和黏着语等。孤立语的特点一般不通过词形变化来表达语法作用，如汉语。屈折语和黏着语的共同点是使用词缀来实现语法功能。但是两者的区别在于屈折语可通过一个词缀实现多个语法功能，而黏着语中的一个词缀一般只具有一个语法功能，因此黏着语中经常会出现一个词内部有多个缀黏着的现象。属于黏着语的语言有日语、韩语、朝鲜语、芬兰语、土耳其语、维吾尔语、蒙古语和哈萨克语等几十种，这些黏着语的特点是词的词汇变化和各种语法变化都是通过在实词词干上连接不同词缀的方式来体现的[19]，因此可以说黏着语是形态丰富的语言。作为黏着语，维吾尔语形态的多变性是维吾尔语最突出的特点之一。

1.2 维吾尔语形态切分的难点

维吾尔语形态切分是维吾尔语自然语言处理的一大难点。导致维吾尔语分词精度不高的原因一般有：黏着性、语音变化现象、歧义和形态切分问题等。

1.2.1 黏着性

维吾尔语作为一种黏着语在语素的组合上具有高度的灵活性，所谓黏着性指的是维吾尔语的绝大部分附加成分都依附在词根之后，在同一个词根上依次连缀几个附加成分，形成一种线条性特点[19]。虽然词干和词缀的数量有限，但是理论上可以组合生成无限的词语，其中，绝大多数维吾尔语词语在语料库中只出现一次[20-21]。维吾尔语通过在词干上添加词缀来实现丰富的句法和语义功能。这种情况在维吾尔语自然语言处理中造成了严重的数据稀疏问题。

1.2.2 语音变化现象

维吾尔语词缀种类多、数目多。在词干和缀、缀与缀连接过程中，由于语音和谐规律，某些词干或词缀会发生弱化、增音、脱落等音变现象。例如，词干باغچا（花园）后面连接后缀لاش（化）后构成新词باغچىلاش（花园化），我们可以发现词干باغچا里的ا弱化为ى。2.4节将在建立好的语料上对语音和谐变化现象进行统计分析。

1.2.3 歧义

维吾尔语词的歧义现象也较严重，这种现象对维吾尔语形态切分任务带来一定的困难。表1给出了一些例子。

表1 维吾尔语的歧义现象举例

维吾尔语词	语义1	语义2
بارماق	手指	去（动词）
توشقان	兔子	装满
ئالما	苹果	别拿
يېڭى	袖子	新

1.2.4 形态切分问题

维吾尔语的形态切分问题还存在意见分歧。传统形态学把形态变化的附加成分分为构词附加成分

（构词词缀）和构形附加成分（构形词缀）。构词词缀的功能是构成新词，构形词缀是不改变词义，而只改变词的语法意义，并表示词的各种语法关系。有的维吾尔语语法书把构词词缀称为“词缀”，而把构形词缀称为“词尾”[22]。上述分类方法有很多不足，还有自相矛盾的地方。维吾尔语里面的一部分附加成分，从形式上看，它们好像是构词附加词缀，但是在功能上它们却具有构形附加成分的功能。例如，بىل+ىم+ىم（我的知识），这里有两个-ىم，形式相同，但功能不一样，第一个-ىم是构词词缀，和بىل-（知道）动词结合构成بىلىم（知识）名词。第二个-ىم是构形词缀，它只是第一人称单数词缀，既不改变بىلىم（知识）的语义，也不改变词性，指名词属于第一人称。因此，对此类词汇进行自动形态切分很难达到预期效果。

2　维吾尔语形态切分标注库建设

2.1　标注规范

2.1.1　基本规则

词干是一个词除去构形附加成分的部分。词干可能由词根构成，也可能由词根加上构词附加成分构成。例如，يازغۇچىلار（作者），-لار是词尾，-ياز是词根，-غۇچى是构词附加成分，这个词除去构形附加成分لار，剩下的يازغۇچى就是词干。

（1）维吾尔语有两种词缀：构词词缀和构形词缀。本文只考虑构形词缀的形态切分，例如，

ساياھەت（旅游）

ساياھەتچى（旅游者）

ساياھەتچى#نىڭ（旅游者＃的）

ساياھەتچىلىك#نىڭ（旅游业＃的）

“旅游者、旅游业”由构词词缀构成，而“旅游者的、旅游业的”由构形词缀构成，本文的形态切分任务是将“旅游者的”和“旅游业的”分别切分成“旅游者＃的”和“旅游业＃的”，而构词成分“旅游者”和“旅游业”不切分。

（2）当词干单独出现时，不加任何标记，默认为词干。例如：旅游。

（3）当词干与构形词缀一起出现时，词干后面“＃”与词缀分开。例如，旅游者＃的。

（4）当词干或词缀发生语音变化时，后面加＄，＄后面写原形。例如，ئائىلى$ئائىلە#لىر$لەر#ى#مۇ。

2.1.2　切分细则

我们主要以名词、形容词、数词、量词、副词、代词、动词为依据来进行切分。目前进行的是粗切分，即构形切分。

（1）名词：名词原形（名词的主格形式）为词干，派生名词（名词的零派生形式）、专用名词可以单独做词干，例如，人名。名词后面加各种名词人称、格、数语法范畴时，名词语法范畴和名词词干分开。

（2）形容词：形容词的原形和最高级被认为是词干（维吾尔语形容词的最高级不带任何构形词缀），减弱和增强级要切分。例如，كۆك#ۇش（浅蓝色）。

（3）数词：数词跟其他成分分开，基数是数词词干，其他形式要切分。如，تۆتىنچى（第四）、ئالتى$ئالتە#مىز（我们六个）等。

（4）量词：量词跟其他成分分开，量词没有加构形附加成分的部分就是量词词干，当量词后面加词缀时词缀和词干要分开。例如，كىلومېتىر#غا（每公里）。

（5）副词：维吾尔语中大部分副词是独立出现的，作为词干来处理，只有极少一部分副词带后缀，这时要将副词与后缀切分开。例如，تېز#راق（快点）、ھازىر#غىچە（直到现在）。

（6）代词：代词单数是代词词干，代词复数要切分。除此之外，维吾尔语代词经常与名词词缀组合，这种形式的代词要与词缀分开。例如，سىز#نى（把你），بىز#نىڭ（我们的）等。

（7）动词：动词带静词化附加成分和时态附加成分。因此动词带的语态附加成分、体语附加成分、否定附加成分、静词化附加成分、时态附加成分、人称附加成分、语气附加成分都与词干切分开。例如，ئال#ىمىز（我们要买）、ياز#دىم（我写了）、ماڭ#دىم（我走了）、تىرىش#ىپ（努力）等。

（8）模拟词：模拟词是词干。

（9）连词：连词单独出现时是词干，附带实词作构形附加成分时要切分。

（10）后置词：后置词是词干。

（11）语气词：单独使用的语气词本身被视为词干，附带实词作构形附加成分的语气词要切分。如，كەل#دى#غۇ（他也来了呢）、سەن#چۇ（你呢）、بەلكىم（可能）等。

（12）感叹词：维吾尔语中的所有感叹词以词干形式出现。

除此之外，维吾尔语中的缩略词基本上存在三种情况。

（1）只取每个词的首字母，并用空格隔开，因此目前不存在切分问题。例如，ج ك پ،ب د ت。

(2) 取第一个词的第一个音节和最后一个词的第一个音节,合并成为一个词干。例如,-پارتكوم等。

(3) 用拉丁字母缩写,作为独立的词。例如,GDP、WTO、KTW 等。

2.2 形态切分语料库建立流程

我们首先从天山网维文版[①]下载了维吾尔语语料,包含新闻、法律、经济和生活等。语料库构建步骤为:爬虫、校对原始语料、分句、校对分句、人工和自动形态切分、人工标注语音和谐变化现象、人工校对形态切分和语音和谐变化现象。语料库包含 10 596 个文档,69 200 个句子,不同领域文档数量的具体分布如表 2 所示。

表 2 不同领域文档数量的领域分布

领域	文档数量
国际	2 817
新疆	1 666
国内	1 651
地区/州	1 555
社会	1 076
乌鲁木齐	877
科教	470
经济	329
其他	157

我们使用 tokenizer. perl(https://github.com/moses-smt/mosesdecoder)工具对语料进行了标点符号切分。同时,为了减轻标注的工作量,我们提取了语料中的词语类型作为人工标注的数据。我们从中央民族大学维吾尔语语言学专业的学生中选择了七位学生对语料进行人工形态切分,要求对每一个词进行带有语音和谐变化的形态切分。在人工标注过程中不断对语料和人工切分错误及不一致性进行更正。人工标注完成后,从七位学生中选出标注最好的一份力克·阿卜杜瓦伊提进行了一次校对,之后又邀请了新疆大学的阿布都热依木·热合曼副教授和这位原标注者交替进行了校对。

2.3 维吾尔语语音变化现象分布

我们对人工切分后的新闻领域语料的词表进行了语音变化现象统计。该新闻语料词表中发生语音和谐变化的词占总词表的 23.9%。为了进一步了解发生语音和谐变化的词中词干和词缀在不同语音和谐变化现象下的分布我们做了进一步统计,统计结果见表 3。

表 3 语音和谐变化现象分布

语素	弱化/%	增音/%	脱落/%
词干	95.7	2.2	2.1
词缀	97.2	1.2	1.6

从表 3 可知,语音变化现象主要体现为弱化,词干和词缀的弱化分布相似。一般情况下,语音和谐变化发生在词干或语素内部,而语素之间不会发生语音和谐变化。由以上分析我们得知维吾尔语中语音和谐变化很严重,而且其中的弱化现象应为研究重点。

2.4 维吾尔语词级形态切分语料库

我们从已进行形态切分的维吾尔语词表(89 923 个)中抽取出一部分建立数据集,用于形态切分任务,该数据集有 19 629 条维吾尔语词。我们将该数据集分为训练集、开发集和测试集。训练集有 17 629 条词,开发集和测试集分别是 1 000 条词。测试任务分为两种:一种是只进行词干和词缀的切分;一种是词干、词缀切分的同时考虑语音变化。该数据集已开源免费使用[②]。目前,已有工作使用该数据集研究了维吾尔语形态切分在神经网络中的性能体现,获得了具有参考价值的实验结果[21],对应的预处理后的数据集和代码也已开源[③]。

2.5 维吾尔语句子级形态切分语料库

我们进一步完善形态切分语料的建设,在词级语料库的基础上建立了句子级形态切分语料。句子级语料包含 69 200 条句子。因为词级形态切分语料建设中已经建立了标注规范,词级规范直接应用到句子中。句子级形态切分时,对句子中的每一个词进行人工形态切分并校对,词干和词缀之间用“#”号来分开,“'”表示右边的语素是词干,“$”表示左边的语素是右边语素的原形。如下例所示:

① http://uy.ts.cn/

② http://thuuymorph.thunlp.org/

③ https://github.com/halidanmu/THUUMS

《 قار ئەجدىها '# دى 》 ناملىق قۇتۇپ رايون '# ى ئىلمىي تەكشۈر '# ۋش پاراخوت '# ى 5- مارپ جۇڭسەن بېكىت '$ بېكەت # ى ئەتراپ '# ى # دىكى پرىز قولتۇق '# ى # دىن يول '# غا چىق '# سپ ، يازلىق ۋەزىپى '$ ۋەزىپە # نى تاماملى # تاماملا # غان 140 نەپەر تەكشۈرگۈچى '# نى ئېل '$ ئال # سپ جۇڭسەن بېكىت '$ بېكەت '# ى '# دىن ئايرى '# ل # دى 。

句子级语料的建设比词级形态语料建设有以下几方面的优势：(1)句子级形态切分时完全可以按上下文来判断句子中每一个词的词干部分，这样就避免兼类词难切分的情况；(2)句子形态切分时可以避免一些正字法、方言词等词汇切分错误。

我们对句子级形态切分语料库进行了统计，统计结果见表 4。通过实验我们发现词、词干、词缀的平均长度是 17、14 和 5，维吾尔语词的最大长度为 33，每个词的词缀的平均个数是 3.5。

表 4　维吾尔语句子级形态切分语料库统计结果

自动统计	字符
词的最长长度	33
词的最短长度	1
词的平均长度	17
词干的最长长度	27
词干的最短长度	1
词干的平均长度	14
缀的最长长度	9
缀的最短长度	1
缀的平均长度	5
词缀的最多个数	6
词缀的最少个数	1
词缀的平均个数	3.5
带词缀的词的比例	0.53
词干发生变形的词的比例	0.12
词缀发生变形的词的比例	0.03
词干与词缀同时变形的比例	0.006
只有 1 个缀变形的比例	0.026
有 2 个或 2 个以上缀变形的比例	0.826 * 10-6e

3　结论

本文描述了构建的维吾尔语形态切分语料库—THUUyMorph，并着重分析了维吾尔语形态切分规则，同时进行了一些语言学上的统计。该语料库已被开源免费使用。该文工作不仅对相关维吾尔语语料库的建设具有参考意义，而且为维吾尔语自然语言处理的研究提供了有益的资源。

参考文献

[1] Afify M, Sarikaya R, Kuo H K J, et al. On the use of morphological analysis for dialectal arabic speech recogni-tion[C]//Interspeech 2006-ICSLP Ninth International Conference on Spoken Language Processing. Pittsbutgh, PA, USA: ISCA, 2006: 277-280.

[2] Botha J A, Blunsom P. Compositional morphology for word representations and language modelling [C]// Proceedings of the 31st International Conference on Machine Learning, Beijing, China: JMLR, 2014: W&CP volume 32.

[3] Clifton A, Sarkar A. Combining morpheme-based machine translation with post-processing morpheme prediction[C]// Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technol-ogies, Portland, Oregon, USA: Association for Computa-tional Linguistics, 2011: 32-42.

[4] Seeker W, Cetinoglu O. A graph-based lattice dependency parser for joint morphological segmentation and syntactic analysis [J]. Transactions of the Association for Computa-tional Linguistics, 2015, 3: 359-373.

[5] Cotterell R, Schutze H. Joint semantic synthesis and morphological analysis of the derived word [J]. Transactions of the Association for Computational Linguistics, 2018, 6: 33-48.

[6] Marslen-Wilson W. Lexical representation and process [M]. Cambridge, MA, USA: MIT Press, 1989.

[7] 哈里旦木·阿布都克里木，刘洋，孙茂松. 神经机器翻译系统在维吾尔语—汉语翻译中的性能对比 [J]. 清华大学学报：(自然科学版)，2017，57(8)：878-883.

[8] Wang L, Cao C, Xia Y, et al. Morphological Segmentation with Window LSTM Neural Networks [C]// Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence. Phoenix, Arizona, USA: Association for the Advancement of Artificial Intelligence, 2016: 2842-2848.

[9] Zohp B, Yuret D, May J, et al. Transfer Learning for

Low-Resource Neural Machine Translation [C]// Proceed-ings of the 2016 Conference on Empirical Methods in Nat-ural Language Processing. Austin, Texas: Association for Computational Linguistics, 2016: 1568-1575.

[10] 吐尔根·依布拉音，阿里甫·库尔班. 基于词典的现代维吾尔语词性自动标注系统的研究[C]. 中国中文信息学会二十五周年学术会议. 北京：中国中文信息学会，2006：148-152.

[11] 艾山·吾买尔. 维吾尔语词法句法分析关键技术的研究[D]. 乌鲁木齐：新疆大学，2009.

[12] 买合木提·买买提，吐尔根·依布拉音. 基于 N-gram 的维吾尔语词性标注研究[C]. 第二届全国少数民族青年自然语言处理学术研讨会. 合肥：中国中文信息学会，2008：206-209.

[13] Yusup A, Lua K T. The development of tagged Uyghur corpus [C]//Proceedings of the 17th Pacific Asia Conference on Language, Information and Computation. Sentosa, Singapore: PACLIC Steering Committee, 2003: 228-234.

[14] Yusup A, Iskender O, and Mamateli T. Progress on con-struction technology of Uyghur knowledge base [C]//Proceedings of the 2009 International Symposium on Intelligent Ubiquitous Computing and Education. Wash-ington, DC, USA: IEEE Computer Society, 2009: 554-557.

[15] Mirejiguli R, Alifu K. Design of the Uyghur FrameNet desktop [J]. Software Engineering, 2015, 3(1): 53-56.

[16] Jiamila W, Wayiti A, Kahaerjiang A, et al. Building con-temporary Uyghur grammatical information dictionary [C]//Proceedings of Worldwide Language Service Infrastructure: Second International Workshop. Kyoto, Japan: Springer Interna-tional Publishing, 2015: 137-144.

[17] Kahaerjiang A, Maihemuti M, and Tuergen Y, et al. Annotation schemes for constructing Uyghur named entity relation corpus [C]//Proceedings of International Conference on Asian Language Processing. Taiwan: IEEE Computer Society, 2017: 103-107.

[18] Cotterell R, Vieira T, Schütze H. A joint model of ortho-graphy and morphological segmentation [C]// Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. San Diego, California: Association for Computiaonal Linguistics, 2016: 664-669.

[19] 艾孜尔古丽，阿力木·木拉提，玉素甫·艾白都拉. 基于形态分析的现代维吾尔语名词词干识别研究 [J]. 中文信息学报，2015，29(6)：208-212.

[20] 哈里旦木·阿布都克里木，程勇，刘洋，等. 基于双向门限递归单元神经网络的维吾尔语形态切分 [J]. 清华大学学报：(自然科学版)，2017，57(1)：1-6.
Abudukelimu Halidanmu, Cheng Yong, Liu Yang, et al. Uyghur morphological segmentation with bidirectional GRU neural networks [J]. J Tsinghua Univ: (SciandTech), 2017, 57(1): 1-6. (in Chinese)

[21] Abudukelimu Halidanmu, Liu Y, Chen X, et al. Learning distributed representations of Uyghur words and morphemes [C]// Proceedings of CCL/NLP-NABD. Guangzhou, China: Springer, 2015: 202-211.

[22] 霍盛. 试论维吾尔语形态变化的功能及其特点 [J]. 新疆大学学报(哲学社会科学版)，1991，(3)：104-111.

哈里旦木·阿布都克里木(1978—)，博士研究生，主要研究领域为自然语言处理。
E-mail：abdklmhldm@gmail.com

刘洋(1979—)，通信作者，博士，副教授，主要研究领域为自然语言处理、机器翻译。
E-mail：liuyang2011@tsinghua.edu.cn

孙茂松(1962—)，博士，教授，主要研究领域为自然语言处理、网络智能、计算社会科学。
E-mail：sms@mail.tsinghua.edu.cn