基于通用依存句法的锡伯语句法树库构建研究

周贺
香港理工大学中文及双语学系
he.zhou@polyu.edu.hk

摘要

我国是一个多民族、多语种的国家，拥有丰富的民族语言资源。然而，使用人口较少、文化影响力较小的语言普遍面临语言濒危的问题，记录和保存这些语言在语言学、民族学与人类学上都具有重要意义。在本研究中，我们以我国仍在活跃使用的满通古斯语——锡伯语为目标语言，从锡伯语语法书、锡伯语报纸《察布查尔报》以及锡伯语《语文》教材中收集了 1200 个句子，以此为语料构建了一个包含词汇、形态以及依存句法信息的树库。本文详细描述了树库的构建过程，深入讨论了标注过程中遇到的难以解决的语言现象，并提出了我们的标注策略。通过标注，我们发现，随着汉语和锡伯语的深层接触，锡伯语不仅在词汇上接受了大量的汉语借词，锡伯语句子结构也受到一定程度的影响。基于所标注的锡伯语树库，我们进行了锡伯语自动句法分析实验，探讨了词、词性、字符特征以及中国少数民族语言预训练模型 CINO 对句法分析性能产生的影响。

关键词： 通用依存句法； 树库； 锡伯语； 满通古斯语族； 低资源语言资源

A Dependency Treebank for Xibe based on Universal Dependencies

He Zhou
Department of Chinese and Bilingual Studies, Hong Kong Polytechnic University
he.zhou@polyu.edu.hk

Abstract

China is a country of multi-ethnics and multi-languages, possessing a wealth of ethnic language resources. However, languages with fewer speakers or less influence are increasingly endangered. Therefore, the documentation of these languages is of significant importance in the fields of linguistics, ethnology and anthropology. This study focuses on the Xibe language, a Tungusic language actively used in China. We collected 1200 written Xibe sentences from a grammar book, Cabcal News and Xibe textbooks, and built a treebank covering annotations on lexical, morphological and syntactic levels. This paper outlines the detailed procedure of treebank construction, discusses challenging cases encountered in annotation, and provides justification for our current annotation decisions. Meanwhile, we also observed that due to extensive language contact with Chinese, modern written Xibe has not only enriched its lexicons from Chinese, but also shows some influence on its syntactic structures. Using the annotated treebank, we conducted dependency parsing experiments, where we investigated the effectiveness of different features including word, part-of-speech, character as well as the Chinese minority pre-trained language model (CINO).

Keywords: Universal Dependencies , Treebank , Xibe , the Manchu-Tungusic Languages , Low-resource Language Resources

第二十三届中国计算语言学大会论文集，第290页-第304页，太原，中国，2024年7月25日至28日。
卷1：主会论文

1 引言

锡伯语 (ISO 639-3:sjo) 是我国锡伯族仍在使用的一门语言，与满语、鄂温克语、鄂伦春语、赫哲语以及历史上的女真语构成我国境内的阿尔泰语系满-通古斯语族语言。根据 2020 年全国人口普查数据 (国务院第七次全国人口普查领导小组办公室, 2021), 锡伯族人口为 19 万 1911 人，占全国人口的 0.01%, 主要分布在东北三省和新疆维吾尔自治区 [0]。其中，使用锡伯语的人口主要集中分布在新疆伊犁哈萨克自治州察布查尔锡伯族自治县及其周边地区，懂锡伯语的人口数在 1 万至 5 万之间 (朝克, 2009; 朝克, 2014)。联合国教科文组织依据语言活力和濒危程度对世界语言进行评估，锡伯语属于严重濒危 (severely endangered) 型语言 (Moseley, 2010)。另外，锡伯语与满语关联密切，两种语言的语法结构和词汇基本一致，学术界多认为锡伯语文是满语文的继续和发展 (安俊, 1985; 赵阿平 et al., 2003)。从文字的角度来看，满通古斯语族的语言中，只有锡伯文字还有一定的使用人口，但是使用人数和范围也在缩小 (朝克, 2014)。因此，从语言学的角度记录和研究锡伯语，不仅对于保护锡伯族的语言和文化具有深远的意义，同时也有助于满语及其文献的研究。

国内外学者对锡伯语的研究主要侧重于利用田野调查收集的语料，描写和分析锡伯语的语音、词汇和语法特点，以及进行跨语言对比研究 (顾松洁, 2016)。然而，采用基于语料库或者计算语言学的方法来记录和分析锡伯语的研究则相对较少。鉴于此，本研究从语言资源建设着手，目标为锡伯语这一低资源语言构建一个包含形态与句法信息的树库。

我们选择使用通用依存句法理论 (de Marneffe et al., 2021, Universal Dependencies, 下文简称为 UD) 作为树库的标注框架。在跨语言或多语言的句法分析任务中，不同语言中的相同句法结构因采用不同的标注框架而呈现出标注差异，这严重影响了句法分析效果。为了消除这一障碍，Nivre et al. (2016) 整合了斯坦福依存句法框架 (de Marneffe and Manning, 2008), 谷歌依存句法框架 (McDonald et al., 2013), 谷歌通用词性标记集 (Petrov et al., 2012) 以及通用形态句法标记集 (Zeman, 2008)，发展出了通用依存句法 (Universal Dependencies)。UD 标注框架自提出以来已广泛应用于世界上多种不同类型的语言。这些开源的树库资源为我们锡伯语树库的标注工作提供了宝贵的参考依据。我们从锡伯语的亲属语言的树库中汲取经验，借鉴其标注方法，从而在确保标注的准确性的同时，也保持了各树库之间标注的一致性。

2 背景介绍

2.1 锡伯语的基本特点

元音 (5)	a[a]	e[ə]	i [i]	o [o]	u[u]
辅音 (19)	n[n]	k [q]/[k]	g [ɢ]/[g]	h [x]/[χ]	b [p]
	p[p^h]	s [s]	š[ʂ]	t [t^h]	d [t]
	l[l]	m[m]	c[tʂ^h]	j[tʂ]	y[j]
	r[r]	f[f]	w[v]	ng[ŋ]	
特定字母（10）	ck[k^h]	cg[g^h]	ch[x^h]	z[z]	ts[ts^h]
	dz[ts]	sy[sz]	tsy[ts^{h}z]	cy[tʂ^{h}ʐ]	jy[tʂz]
满语元音 (1)	v [u]				

表 1: 锡伯语字母、其拉丁字母转写及国际音标

锡伯语与阿尔泰语系内其他亲属语言，如蒙古语族和突厥语族语言，在语言的多个层面均呈现出了共同的特征。在语音层面，锡伯语存在元音和谐现象，即一个词内的元音一般具有相同的发音特征；在形态层面，锡伯语采用黏着形态，通过在词根上附着不同的词尾来表达语法功能，而且锡伯语的黏着词缀与词根同样遵循元音和谐规律；在句法层面，锡伯语句子遵循主

[0] https://www.neac.gov.cn/seac/ztzl/xbz/gk.shtml

语料来源	句子数	总词元数	句长 $_{max}$	句长 $_{min}$	平均句长	中位数	方差 σ
语法例句	544	5 757	50	3	10.58	10	5.43
《察布查尔报》	266	9 644	90	3	36.26	33	21.52
小学课本	390	8 062	110	3	20.67	17	13.89
总计	1 200	23 463	110	3	19.55	13.5	16.65

表 2: 锡伯语树库语料句子的统计信息（句长的计算基于用空格分词处理之后的词元数）

语-宾语-谓语（SOV）的语序，短语遵循中心词后置（head-final）的语序。现代锡伯文文字是在圈点满文的基础上稍作修改演化而来，采用从上到下、从左到右的书写顺序。表1为锡伯语书面语字母及其拉丁字母转写和国际音标，锡伯语有 5 个元音，19 个辅音，还有 10 个用于转写外来词的特定字母以及满语元音 ᡶ（*v*）[1]。1947 年“锡索文化协会”在锡伯文字改革时，删除了满语元音 ᡶ（*v*）(安双成, 1997)，但是由于锡伯语继承了众多满语词汇，这个字母依然广泛地存在于现代锡伯语文本中。

与我国仍存的其他满通古斯语族语言 (满语、鄂温克语、鄂伦春语和赫哲语) 相比，现代锡伯语不仅同时存在书面语和口语，而且掌握两种形式的人数也多于其他语言。然而，锡伯语的书面语和口语在语音、词汇以及语法上都有较大差别 (苏承志, 1995; 佟加·庆夫, 1996; 张泰镐, 2008)。语音上，书面语词汇在口语中经历了元音音变或脱落、音节脱落以及辅音异化等变化，书面语中的多音节词在口语中逐渐双音节或单音节化；词汇上，书面语多使用规范词汇，而口语中并未使用，例如锡伯语口语中的话题标记 *da* 在书面语中并未得到使用 (Jang, 2020)；语法上，口语句子短小多变，结构简单，而书面语语法规范、准确且完整 (顾松洁, 2016)。在本研究中，我们选择锡伯语书面语作为树库的语料，主要出于以下三个原因：首先，虽然现存锡伯语口语在整体上不存在方言的区分，但新疆地区的锡伯语仍展现出细微的地区差异。国内外的研究中所使用的口语语料，例如张泰镐 (2008), Zikmundová (2013)，大多采用拉丁字母或者国际音标转写，缺少统一的文字表示，这不利于树库语料的收集和整理。其次，书面语句子严格遵守句法规则，结构规整，有助于我们准确地分析和理解锡伯语的词汇、形态及句法结构。最后，以锡伯语书面语为媒介的文学作品、语言课本和《察布查尔报》等丰富的语言资源，为语料库的构建提供了宝贵的语料。

2.2 锡伯语和满语的关系

锡伯族的固有语言与满语相近，锡伯人在被纳入八旗后，很快学用了满文并接受了逐步规范化了的满语。18 世纪中期，清朝为了加强西北边陲的防务，从东北抽调了一批锡伯兵西迁至新疆伊犁河谷戍边屯垦。这些锡伯人及其后代形成了今天新疆伊犁的锡伯族。在漫长的历史进程中，锡伯族与周边其他民族接触较少，居住较为集中，而且受到封建封闭式的八旗制度的约束，所以西迁的锡伯人很好地保留了满语文。至 20 世纪 40 年代，伊犁的锡伯族在旧满文的基础上进行修订和改革，形成了现代锡伯文 (佘吐肯, 2006)。从历史语言学的角度，锡伯语和满语在文字、语音、词汇和语法方面均存在着高度的一致性。现代锡伯文字是在圈点满文的基础上稍作修改演变而来的；在满语和锡伯语的基本词汇中，同源词达到 90% 以上，而且二者的语法结构也基本一致 (朝克, 2014)。国内外学者对于锡伯语的独立地位持有不同的看法，本文对此不做深入探究，主要采用佘吐肯 (2006) 的观点，认为锡伯语是一门独立的语言，它与满文具有源流的关系，在继承满文的基础上，继续丰富和发展了满文。基于此，在树库标注的过程中，除了锡伯语的参考资料，我们还可以参考满文语言材料，如《新满汉大辞典》(胡增益, 2020)，Gorelova 编写的满文语法 *Manchu Grammar* (Gorelova, 2002)，Jerry Norman 编著的满-英词典 *A Comprehensive Manchu-English Dictionary* (Norman, 2020) 以及 Roth Li 编写的满文读本 *Manchu-A Textbook for Reading Documents* (Li, 2000) 等。

```
# sent_id = grammarbook_sjo_p1_9
# text = ᠮᠢᠨᠢ ᠠᠮᠠ ᠤᠷᠤᠮᠴᠢ ᠴᠢ ᠶᠠᠪᠤᠬᠠ ᠃
# text[phon] = mini ama urumci ci yabuha .
# text[chn] = 我的爸爸去了乌鲁木齐。
1  ᠮᠢᠨᠢ  ᠮᠢᠨᠢ  PRON   _  _                        2  nmod:poss  _  Translit=mini
2  ᠠᠮᠠ   ᠠᠮᠠ   NOUN   _  _                        5  nsubj      _  Translit=ama
3  ᠤᠷᠤᠮᠴᠢ ᠤᠷᠤᠮᠴᠢ PROPN  _  _                        5  obl        _  Translit=urumci
4  ᠴᠢ    ᠴᠢ    ADP    _  Case=Abl                   3  case       _  Translit=ci
5  ᠶᠠᠪᠤᠬᠠ ᠶᠠᠪᠤ   VERB   _  Tense=Past|VerbForm=Fin  0  root       _  Translit=yabuha
6  ᠃     ᠃     PUNCT  _  _                        5  punct      _  Translit=.
```

图 1: 用 CoNLL-U 格式表示的依存结构例句

3 数据收集及预处理

3.1 语料收集

我们为锡伯语树库共收集了 1200 个句子，其中从《锡伯语语法通论》(佘吐肯, 2009) 收集了 544 个语法例句（共计 5757 个词元），从《察布查尔报》2019 年 1 月至 3 月的 9 期报纸中收集了 266 句新闻语料（共计 9644 个词元），另外从锡伯文小学课本《ᠨᠢᠶᠠᠮᠠᠩᠭᠠ ᠭᠢᠰᠤᠨ》(*niyamangga gisun*, “语文”) 第 3 册至第 6 册 (何文勤, 2006) 收集了 390 句（共计 8062 个词元）。我们统计了每一部分句子的句子数和句长等信息，如表2所示。在语料收集过程中，考虑到诗歌和文学语体的特殊性，我们在这一版本的树库中暂未收录。

3.2 数据预处理及标注

在 UD 中，每一个句子都用 CoNLL-U 格式 (Buchholz and Marsi, 2006) 来表示（见图1)。一个 CoNLL-U 格式的句子包括三类信息：(1) 以 # 标记开头的行存储句子的元信息，用来记录句子在树库里的编号（sent_id)，锡伯语文本（text)，句子的拉丁转写（text[phon]）以及句子的中文或英文翻译。(2) 句子以空格切分成词元（token)，每一个词元标注其词源编号（ID)，词形（FORM)，原形（LEMMA)，通用词性（UPOS,universal part-of-speech)，扩展词性（XPOS,extended part-of-speech)，形态特征（FEATS)，依存词（HEAD)，依存关系（DEPREL)，增强依存关系（DEPS）以及其他自定义信息（MISC)。10 个字段之间用制表符分隔开，无值字段用下划线表示。(3) 句子和句子之间的空行。

我们自动获得每一个锡伯语句子的元信息。锡伯文文本首先根据表1通过 python 脚本自动转写成拉丁形式，然后将其连同其拉丁转写及中英文翻译转换为 CoNLL-U 格式。对于每一个词元的信息，我们采用手工标注的方式。标注由两名标注者完成，其中第一名标注者完成了 464 句语法例句、新闻和小学教材的全部句子，第二名标注者完成了 80 句语法例句。由于两名标注者均为语言学专业的非锡伯语母语者，为了确保标注质量，标注和校对的过程中标注者不仅参考了锡伯文满文语法，还通过咨询锡伯语母语者获得句法知识。另外，标注工作借助 UD Annotatrix (Tyers et al., 2017) 和 conllueditor (Heinecke, 2019) 工具来提高标注效率。标注完成后，我们统计了锡伯语树库中所涉及的词性、形态和依存关系，详细数据列于表B.1、表C.1和表D.1中。

图1为一个标注完成的实例，其中对每一个词元，我们标注其原形、通用词性、形态信息，为方便阅读，我们将每一个词的拉丁转写存储在每一行的最后一列。例如，这个句子的谓语动词 ᠶᠠᠪᠤᠬᠠ（*yabuha*）的索引号为 5，是动词 ᠶᠠᠪᠤ（*yabumbi*，“行走”）的一般过去时形式，因此它的词性为 VERB, 形态特征标注为过去时（Tense=Past）限定动词（VerbForm=Fin)。作为句子的谓语，它的依存词为索引号为 0 的根节点，依存关系为 root。另外，最后一列为该词的拉丁转写（Translit=yabuha)。

4 树库标注中的难点

在树库标注的过程中，UD 的标注规范基本上可以覆盖锡伯语的形态和句法特征。但是，如第2节所提到的，现代锡伯语在继承了满语的基础上继续丰富和发展，随着锡伯族与其他民族的

[1]为了排版方便，正文中的锡伯语文本水平显示。

分写形式	ᠪᠣ᠊ᠤ᠊ᠭᠤᠸᠠᠨ᠊	ᠫᠠᠨ᠊ᠭᠤ᠊	ᠪᠢᠶᠠᠨ᠊ᠫᠣᠣ᠊	ᠭᠤᠩ᠊ᠬᠡ᠊ᠴᠭᠣ᠊	ᠴᠢ᠊ᠶᠡ᠊
连写形式	ᠪᠤᠤ᠊ᠭᠤᠸᠠᠨ᠊	ᠫᠠᠨ᠊ᠭᠤ᠊	ᠪᠢᠶᠠᠨ᠊ᠫᠣᠣ᠊	ᠭᠤᠩ᠊ᠬᠡ᠊ᠴᠭᠣ᠊	ᠴᠢ᠊ᠶᠡ᠊
拉丁转写	bo u guwan	pan gu	biyan poo	gung he cgo	ci ye
中文	“博物馆”	“盘古”	“鞭炮”	“共和国”	“企业”

表 3: 锡伯语汉语借词示例

1-3	ᠪᠤ᠊ᠭᠤᠸᠠᠨ᠊								
1	ᠪᠤ	ᠪᠤ	X	_	_	10	nsubj	_	Translit=bo
2	ᠤ᠊	ᠤ᠊	X	_	_	1	flat	_	Translit=u
3	ᠭᠤᠸᠠᠨ᠊	ᠭᠤᠸᠠᠨ᠊	X	_	_	1	flat	_	Translit=guwan

图 2: 多词元词的标注示例，以 ᠪᠤ᠊ᠭᠤᠸᠠᠨ᠊, *bo u guwan*, “博物馆”为例

交流日益频繁，锡伯语与汉语、维吾尔语和哈萨克语等语言的接触也逐渐增多。语言接触的一个突出表现是，锡伯语从汉语中借来了大量的词汇，同时，一些句子结构也在汉语翻译的影响下发生了细微的变化。我们选取标注中遇到的特殊语言现象，在详细分析的基础上决定标注方法。

4.1 由汉语借词产生的多词元词

根据 UD 标注规范，依存关系应建立在词与词之间，而这里的“词”是指句法意义上的词，并非正字法意义上的词。例如，法语的 au 是介词 à 和阳性冠词 le 的缩合形式，au 在正字法意义上算作一个词，然而从句法的角度应该算作两个词，因此在 UD 标注中，这类缩合冠词被拆分为两个词，au 被当做一个多词词元（Multi-word Token）。与此相反，锡伯语中出现了需要将几个词元合并为一个词的特殊现象，这类词以汉语借词为主，我们称之为多词元词（Multi-token Word）。

表3列出了 5 个汉语借词，锡伯语从汉语借词的方式是将汉语音节直接转写为发音相同或相似的一个或多个锡伯语音节，书写时每一个汉语音节对应的锡伯语音节写成锡伯语的一个“词”。在遵循锡伯文正字法 (新疆维吾尔自治区民族语言文学工作委员会, 1992) 的前提下，多词元词也可以连写，例如 ᠫᠠᠨ᠊ᠭᠤ（*Pan Gu*,“盘古”）也可以写作 ᠫᠠᠨᠭᠤ。然而，ᠪᠤ᠊ᠭᠤᠸᠠᠨ᠊（*bo u guwan*，“博物馆”）只有分写形式而不能连写，因为锡伯文中元音 ᠣ（*o*）与 ᠤ（*u*）没有连写形式。

在语料预处理阶段（见第3.2节），锡伯语句子按照空格切分成了词元串。词元串中既有锡伯语的固有词，也有汉语借词的不成词语素。图2为 ᠪᠤ᠊ᠭᠤᠸᠠᠨ᠊(*bo u guwan*) 的切分和标注。在标注这类多词元词时，我们在第一个词元前插入一行，标注多词元词的前后边界（1-3）和词形，其余字段为空。由于构成该词的三个词元在锡伯语中均不能单独成词，我们无法确定它们的具体词性类别，因此它们的词性确定为 X。根据 UD 对外来词或短语的标注方法，我们把多词元词的内部结构标注为一个平面结构，将第一个词元作为中心词，其后的两个词元分别依存于第一个词，依存关系为 flat。这样的标注方式取消了多词元词内部的层级结构，不会与短语结构相混淆，也不会影响句法层级结构。

4.2 话题标记 ᠣᠴᠢ*oci* 的特殊用法

在锡伯语书面语中，论元后一般用“格”来表示其句法功能，主格为零标记。但是，名词谓语句中的主语和名词性谓语之间有时会出现话题标记。ᠣᠴᠢ（*oci*）是锡伯语书面语中话题标记之一，它的基本功能是标记一个句子的主题。例如，在图3中，ᠣᠴᠢ（*oci*）位于主语 ᠪᠧᠵᠢᠩ（*beijing*,“北京”）之后，名词短语 ᠮᠤᠰᠧ ᠭᠤᠷᠤᠨ ᠢ ᠭᠧᠮᠤᠨ ᠬᠧᠴᠧ（*musei gurun i gemun hecen*，“我们国家的首都”）之前，其功能是用来标记话题。母语者经常将 ᠣᠴᠢ（*oci*）等同于现代汉语系动词“是”。然而在句法功能上，锡伯语 ᠣᠴᠢ（*oci*）与韩语的은/는（*eun/neun*）、日语的は（*wa*）更相近，与现代汉语中的“是”并不对应。因为“话题”是语用学概念，而且并非所有的语言都是话题优先型语言，UD 的表层标注中并未包含对句子信息结构的标注。因此，我们参考 UD 中其它话题优先型语言的标注方法，如韩语和日语，将话题标记当做特殊的主格标记，ᠣᠴᠢ（*oci*）的词性为后置词（ADP），

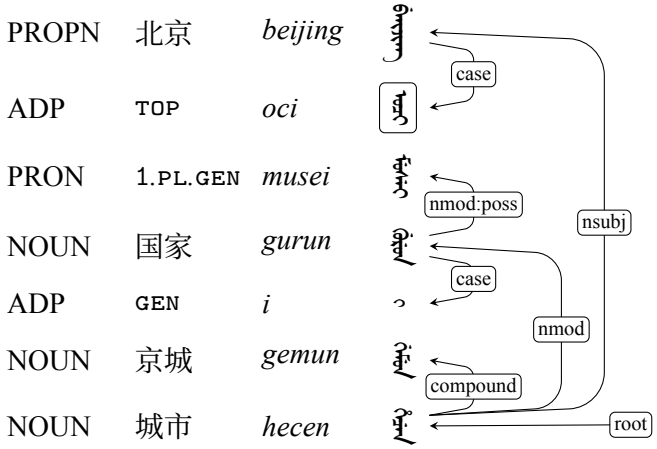

图 3: 依存结构树“北京是我们国家的首都”

它依存于主语，二者之间的依存关系为 case。

(1) a. **是**有利于群众的事他都肯干。（取自《现代汉语词典》）

b. uranda-ra jilgan be [*pro* oci šan]$_{clause}$ teni donji-mbi
发响-PTCP.IPFV 声音 ACC [*pro* COP 耳朵]$_{clause}$ 就 听见-PRS
“发出的声响是耳朵就能听见” [2]

现代汉语中“是”除了做系动词之外，还可以用在名词前，表示“凡是”的意思，如例 (1a)。由于锡伯语母语者经常将 ᠣᠴᡳ（*oci*）直接等同于现代汉语“是”，“是”的这一功能也被锡伯语借用。例 (1b) 取自第五册《ᡧᠠᠨ ᠪᡝ ᡤᡳᡩᠠᠮᡝ ᡥᡠᠨᡤᡤᠣᠨ ᡥᡡᠯᡥᠠᠮᠪᡳ》（*šan be gidame honggon hvlhambi*，“掩耳盗铃”）。这句话中 ᠣᠴᡳ（*oci*）位于 ᡧᠠᠨ（*šan*，“耳朵”）之前，意为“凡是耳朵就可以听到发出的声响”。ᠣᠴᡳ ᡧᠠᠨ（*oci šan*）位于句子的宾语和谓语动词之间，我们将其分析为一个嵌入在主句中的主语脱落（*pro*-drop）的小句。这样，ᠣᠴᡳ（*oci*）的功能是做小句内的系动词。

图4是该句的依存结构分析，ᠣᠴᡳ（*oci*）依存于 ᡧᠠᠨ（*šan*，“耳朵”），二者的依存关系为 cop（系动关系）。小句依存于句子的谓语动词 ᡩᠣᠨᡤᡳᠮᠪᡳ（*donjimbi*，“听到”），是主句的副词性从句（advcl）。由于采用了从汉语到锡伯语逐词翻译的方式，类似例（1b）的句子忽视了锡伯语的自然表达习惯和句法结构。这不仅反映了翻译过程中的一些问题，同时也揭示了汉语对锡伯语在一定程度上的影响。

4.3 ᠨᡳᠩᡤᡝ *ningge* 的两种用法

ᠨᡳᠩᡤᡝ（*ningge*）是属格标记ᡳ（*i*）/ᠨᡳ（*ni*）和名词化词缀ᡤᡝ（*-ngge*）相结合而形成的。在我们收集的语料中，ᠨᡳᠩᡤᡝ（*ningge*）主要出现在两种句法环境中。

(2) a. ere bithe oci sefu i bithe .
这 书 TOP 老师 GEN 书 .
“这本书是老师的书。”

b. ere bithe oci sefu i **ningge** .
这 书 TOP 老师 GEN PTCL.POSS .
“这本书是老师的。”

c. ere bithe oci **miningge** .
这 书 TOP 1SG.GEN.POSS .

[2]语法注释释义详见表A.1

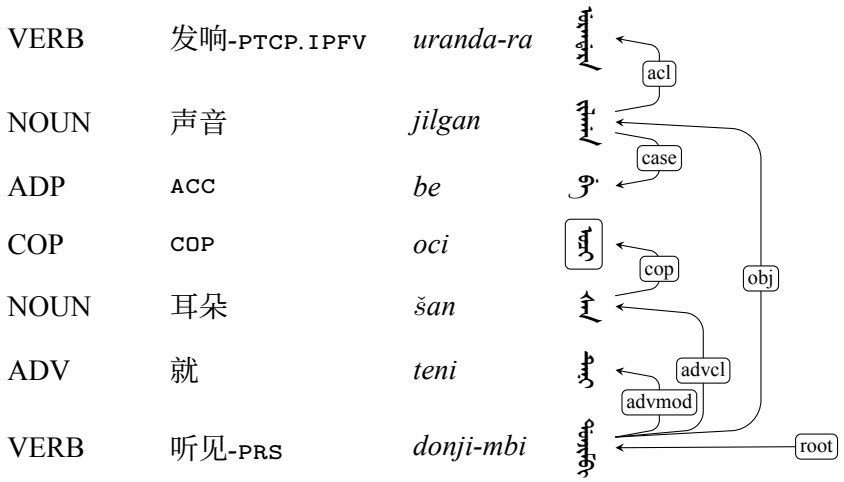

图 4: 依存结构树“发出的声响是耳朵就能听见”

“这本书是我的。”

第一种情况，ᠨᠢᠩᠭᠡ（*ningge*）附加名词或代词后，做表示领属关系的助词，替代由名词/代词和中心语所组成的名词短语 (胡增益, 2020)。例（2a）是一个名词谓语句，谓语是一个表领属关系的名词短语 ᠰᠡᠭᠦ ᠶᠢ ᠪᠢᠲᠡ（*sefu i bithe*，“老师的书”），锡伯语用属格ᠶ（*i*）标记“老师”和“书”之间的领属关系。（2b）是（2a）的省略形式，因为主语中已经提到过“书”，有时为了避免重复，中心词再次提及时可以省略，用 ᠨᠢᠩᠭᠡ（*ningge*）做领属关系助词来补足省略的成分，与前词构成表领属关系的名词短语的省略形式，语义上表示抽象概念的“东西”。例（2b）中，中心语 ᠪᠢᠲᠡ（*bithe*，“书”）省略，ᠨᠢᠩᠭᠡ（*ningge*）是属格ᠶ（*i*）和名词化词缀 ᠩᠭᠡ（*-ngge*）结合形式，与 ᠰᠡᠭᠦ（*sefu*，“老师”）一起构成表领属关系的名词短语，ᠰᠡᠭᠦ ᠨᠢᠩᠭᠡ（*sefu ningge*）意为“老师的（东西）”。当领属性名词短语的修饰语是人称代词，人称代词采用属格形式，例如例 (2c)，ᠨᠢᠩᠭᠡ（*ningge*）附加在形容词性物主代词之后，将整个短语变为名词性物主代词 (佘吐肯, 2009)。

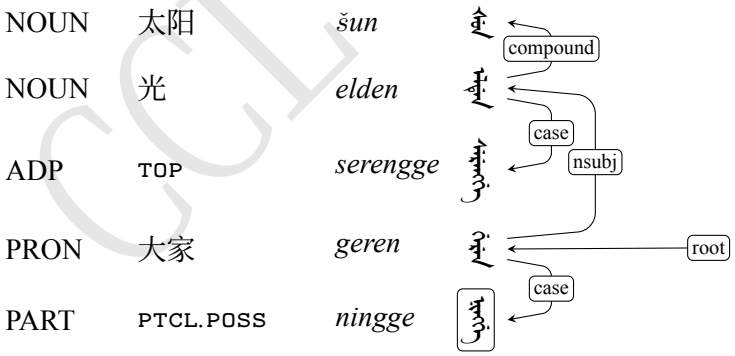

图 5: 依存结构树“阳光是大家的”

图5例句取自锡伯《语文》第六册《ᠰᡠᠨ ᡝᠯᡩᡝᠨ》(*šun elden*，“阳光”)，句中名词性谓语 ᠭᡝᠷᡝᠨ ᠨᠢᠩᠭᠡ（*geren ningge*，“大家的”）可以理解为“大家的阳光”或“大家的东西”，ᠨᠢᠩᠭᠡ（*ningge*）的作用是补足省略的中心语。在标注中，ᠨᠢᠩᠭᠡ（*ningge*）的词性为 PART（助词）, 它依存于前面紧邻的名词或代词，依存关系为 case，表明该助词赋予前面名词领属格。这种标注方式将 ᠭᡝᠷᡝᠨ（*geren*，“大家”）标注为名词短语的中心词，从语义上看不尽合理，主要是为了遵守 UD 的标注规则。UD 标注体系为了更好地捕捉不同语言之间平行性和相似句法结构，规定中心词一般为实词，虚词只能依存于实词。

另外，锡伯语书面语中，ᠨᠢᠩᠭᠡ（*ningge*）还常位于做主语的名词短语或形动词短语之后，做体词化标记（substantivizer） (Gorelova, 2002)，提示句子的话题。如图6所示的一个取自《察布

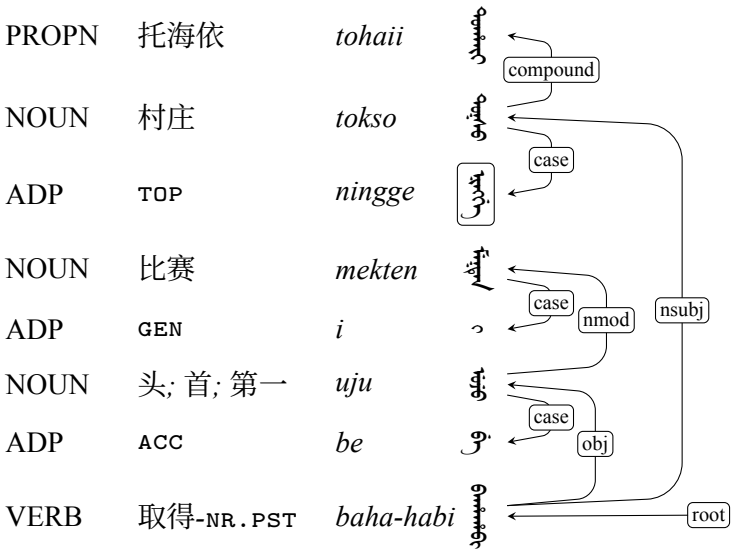

图 6: 依存结构树“托海依村获得了比赛的第一名”

查尔报》的例句，句中 ᠨᠢᠩᠭᠡ（*ningge*）在主语 ᠲᠣᠬᠠᠶ ᠲᠣᠭᠰᠣ（*tohaii tokso*，“托海依村”）之后，从句法和语义的角度来分析，我们将其分析为话题标记，它的词性为后置词（ADP），依存于前面的主语，二者依存关系为 case，表明该词赋予前面名词短语特殊的主格（同第4.2节 ᠤᠴᠢ（*oci*）的标注方法）。

4.4 无中心语的关系子句

关系子句（Relative Clause）的句法功能是做名词的修饰语。不同的语言中，关系子句和中心词的相对位置有前后之分，例如，英语的关系子句由关系代词引导，位于中心词之后；汉语的关系子句位于中心词之前，二者之间用“的”字连接。根据中心词在关系子句中可充当的成分，关系子句可以分为有空缺的关系子句和无空缺的关系子句，有空缺的关系子句又包括主语位置空缺和宾语位置空缺两类。例如，(3a) 中，中心语位置上的“人”是关系子句谓语“看见”的主语，而关系子句的主语位置为空，因此被称为主语空缺的关系子句；类似地，(3b) 为宾语位置空缺的关系子句。无空缺的关系子句是指中心词与关系子句中的主语和宾语均无对应关系，如 (3c)，关系子句不缺少主语或宾语，中心词“时间”也不与之相对应。在 UD 的标注体系中，关系子句的谓语动词依存于关系子句修饰的中心词。在无空缺的关系子句中，从句的谓语与中心词之间的依存关系使用 acl（修饰名词的从句）来标记；而在有空缺的关系子句中，则用 acl:relcl 来特殊标记。

(3) a. 主语空缺的关系子句：[昨天 t_i 看见他的]$_{RC}$ 人 $_i$
b. 宾语空缺的关系子句：[昨天他看见 t_i 的]$_{RC}$ 人 $_i$
c. 无空缺的关系子句：[我写文章的]$_{RC}$ 时间

锡伯语关系子句置于中心词之前，从句中的谓语动词采用形动词形式，而且从句的主语由属格标记，以与主句的主语进行区分。锡伯语形动词由动词词根附加形动词词尾构成，形动词词尾有完成体（4a）、非完成体（4b）以及进行体（4c）之分，其中完成体和非完成体后缀的选择基于动词词根由元音和谐律决定。

(4) a. 完成体后缀：ᠬᠠ（*-ha*），ᠬᠡ（*-he*），ᠬᠣ（-ho）
b. 非完成体后缀：ᠷᠠ（*-ra*），ᠷᠡ（*-re*），ᠷᠣ（*-ro*）或 ᠺᠠ（*-ka*），ᠺᠡ（*-ke*），ᠺᠣ（*-ko*）
c. 进行体后缀：ᠮᠠᠬᠠ（*-maha*）

图7例句中，“我爸爸种的玉米”是被关系子句修饰的名词短语，从句的主语是 ᠮᠢᠨᠢ ᠠᠮᠠ（*mini ama*，“我的爸爸”），后有属格标记；从句的谓语 ᠲᠠᠷᠢᠬᠠ（*tariha*，“种植”）是在词根 ᠲᠠᠷᠢ（*tari-*）上

附着完成体形动词词缀 ᠬᠠ（*-ha*）构成的形动词，修饰中心语 ᠪᠣᠯᠢᠮᠣ（*bolimo*，“玉米”）。从句内，ᠲᠠᠷᠢᠬᠠ（*tariha*，“种植”）前的宾语位置为空，中心语 ᠪᠣᠯᠢᠮᠣ（*bolimo*，“玉米”）是其语义上的宾语，因此例7符合宾语空缺的关系子句的界定。因此，根据标注规范，ᠲᠠᠷᠢᠬᠠ（*tariha*，“种植”）依存于 ᠪᠣᠯᠢᠮᠣ（*bolimo*，“玉米”），二者之间有 acl:relcl 的依存关系。

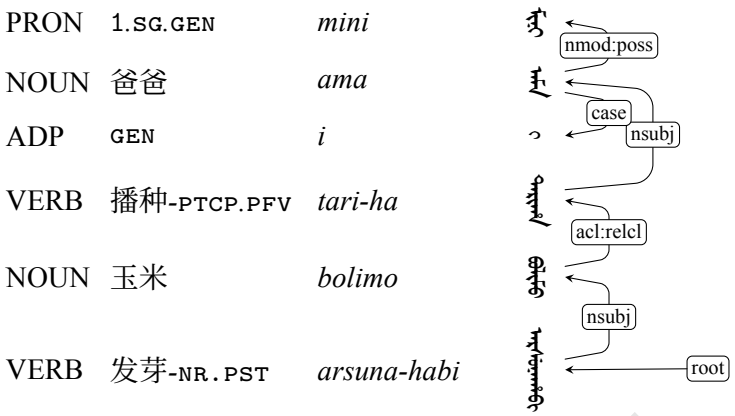

图 7: 依存结构树“我爸爸种的玉米发芽了”

除此之外，锡伯语还存在一种较为特殊的无中心词的关系子句。关系子句所修饰的中心词被，但是关系子句的谓语动词词形需要在形动词基础上再加上名词化词尾 ᠩᠭᠡ（*-ngge*），变成名动词。该词尾将形动词体词化，可以表达：（1）动作的抽象概念；（2）具有物质或非物质性质的客体；（3）行为的主体 (Gorelova, 2002)。图8所示例句中，ᠪᠡᠶᠡ ᠸᠠᠯᠢᠶᠠᠲᠠᠢ ᠬᠤᠯᠬᠠ ᠪᠡ ᠵᠠᠪᠠᠬᠠᠩᠭᠡ（*beye waliyatai hvlha be jafahangge*，“舍身擒拿盗贼的”）是一个省略了中心词 ᠨᠢᠶᠠᠯᠮᠠ（*niyalma*，“人”）的关系子句，此处名词化词尾 ᠩᠭᠡ（*-ngge*）表示“抓”的行为主体。所以，在句法层面，这种关系子句没有显性的中心词，但是在语义层面中心词隐性地存在。在标注中，这类关系子句没有用到表示关系子句的依存关系标记，名词化了的关系子句充当句子的主语。

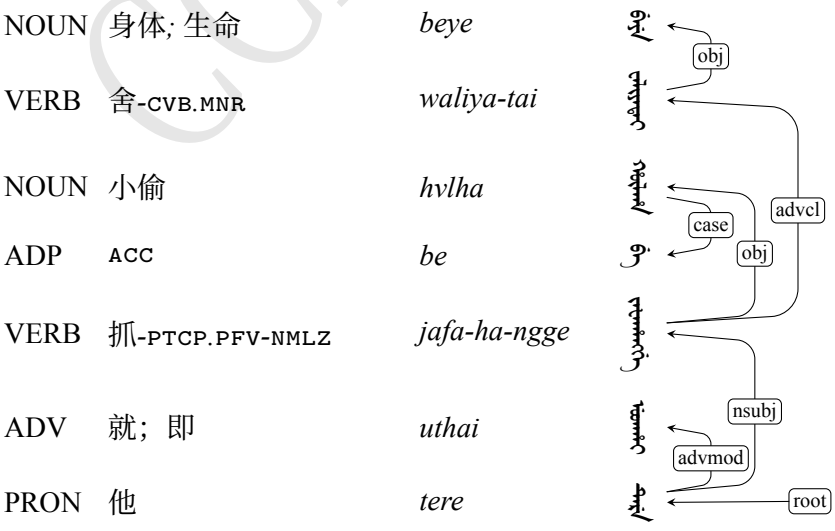

图 8: 依存结构树“舍身擒拿盗贼的就是他”

5 句法分析实验

5.1 数据切分及实验设置

由于树库的规模较小，为了可以更加全面地评估句法分析的性能，我们采用三折交叉验证的方法。如第3节所述，锡伯语树库的句子包括语法书例句、新闻与语文课本句子，且 3 种类型的语料数量并不均衡。为了保证每一折语料分布的均衡性，我们使用分层抽样的方法将句子按照语料体裁平均分成三折，每折 400 句。在实验中，每一折都用作训练、验证和测试一次，从而使每一个句子都可以参与一次训练和测试。

我们使用基于双仿射依存句法分析 (biaffine dependency parsing) 模型的句法分析器 MultiParser[3] (Sayyed and Dakota, 2021)。为了探究词性和字符信息是否可以有效地提高句法分析的准确率，除了词向量，我们还加入词性和字符向量来增强词的表示。此外，我们还使用从少数民族多语言模型 CINO (Yang et al., 2022, CINO-large-v2) 提取的词向量。

在测试阶段，我们计算 UAS(Unlabeled Attachment Score) 和 LAS(Labeled Attachment Score)。UAS 计算参与测试的句子中找到正确的支配词 (head) 的词所占的比例，LAS 计算测试集中既找到正确的支配词又正确地标记了二者之间依存关系 (dependency relation) 的词所占的比例。在本文中，我们计算三折交叉实验中得到的 UAS 和 LAS 的均值。

5.2 实验结果及分析

特征	UAS	LAS
`word`	68.19	56.57
`word+pos`	67.19	55.52
`word+char`	71.65	62.34
`word+pos+char`	71.69	62.17
`word+char+CINO`	**72.11**	**62.47**

表 4: 使用不同特征输入表示的句法分析结果。其中，`word` 指词向量, `pos` 指词性向量, `char` 指字符向量。

表4为使用不同的特征组合的条件下句法分析的准确率。当仅用词向量作为输入时，句法分析获得 68.19% 的 UAS 和 56.57% 的 LAS，但是当加入词性向量时，句法分析性能下降，LAS 下降了 1.05%。究其原因，由于训练数据较少，测试数据中平均高达 46.38% 的词在训练数据中并未出现，因而模型并未充分地学习到词汇和句法信息。当我们继续加入词性信息时，词性信息并未帮助消歧，反而增加了噪音，使句法分析表现下降。当输入表示用词向量和字符向量时，LAS 提高了 5.77%(62.34 vs. 56.57)，这表明字符向量提供给句法分析有用的词汇信息以及字符之间的位置信息，有效地缓解了未登录词的问题。然而，当输入表示将词、词性和字符向量相结合时，句法分析性能并没有进一步提高，LAS 较使用词和字符向量的模型下降了 0.17%，虽然这一差异并不显著，这仍然可以表明词性信息在本实验设置条件下并未起到正面的作用。

当使用词、字符和 CINO 向量作为输入时，句法分析性能达到最佳，UAS 达到 72.11%，LAS 达到 62.47%。但是，该模型只高于使用词和字符特征的模型 0.13%(62.47 vs. 62.34)，这表明 CINO 语言模型能提供给锡伯语句法分析的信息也较为有限。CINO 模型基于多语言训练模型 XLM-R (Conneau et al., 2020)，在藏语、蒙语、维吾尔语、哈萨克语、朝鲜语、壮语和粤语上进行二次训练，其训练数据中并没有锡伯语数据，与锡伯语文字相似的是回鹘体蒙文。我们进一步对比了 Unicode 编码表[4]中蒙古文和锡伯文字母的差异，发现锡伯语只有 12 个字母与蒙古文具有相同的 unicode 编码，其余 21 个字母为锡伯语独有，是 CINO 预训练语言中所不覆盖的。因此，锡伯语句法分析性能得到轻微提升，主要归功于 CINO 模型中与蒙古语同码的 12 个字母所带来的知识。

[3] https://github.com/zeeshansayyed/multiparser
[4] https://www.unicode.org/charts/PDF/U1800.pdf

6 小结及下一步工作

在本研究中，我们基于通用依存句法，为我国资源稀缺的少数民族语言——锡伯语，手工标注了一个含有词汇、形态和句法信息的依存树库。树库目前包括从《锡伯语语法通论》、《察布查尔报》和锡伯《语文》收集的 1200 个标注完整的句子。我们详细介绍了树库的标注过程，深入分析了标注过程中遇到的较难处理的词汇和句法层面的语言现象，并提出了我们的标注策略。这些语言现象同时也反映出，与汉语的深层次接触导致锡伯语不仅在词汇层面上产生变化，锡伯语句法也产生了微妙的变化。基于所标注的锡伯语树库，我们进行了锡伯语句法分析实验，实验中我们重点研究了不同的特征对句法分析器性能的影响，发现在该实验条件下，输入字符可以为句法分析提供有用的信息，而词性会带来干扰。另外，受预训练数据和文字类型所限，锡伯语句法分析器从少数民族预训练语言模型 CINO 中受益有限。

我们的下一步工作主要包括两个方面：首先，我们将继续从《察布查尔报》和锡伯语文学作品中收集语料，扩大树库的规模；其次，手工标注句法树库是一项耗时耗力的工作，我们将以手工标注的树库为基础，继续深入研究锡伯语自动词法分析和句法分析，实现树库的自动标注。与此同时，我们也将扩大工作范围，为其他的少数民族语言构建语言资源及应用。

参考文献

Sabine Buchholz and Erwin Marsi. 2006. Conll-x shared task on multilingual dependency parsing. In *Proceedings of the tenth conference on computational natural language learning (CoNLL-X)*, pages 149–164.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Oline.

Marie-Catherine de Marneffe and Christopher Manning. 2008. The stanford typed dependencies representation. In *Coling 2008: proceedings of the workshop on cross-framework and cross-domain parser evaluation*, pages 1–8.

Marie-Catherine de Marneffe, Christopher Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal dependencies. *Computational Linguistics*, 47(2):255–308, 07.

Liliya M Gorelova. 2002. *Manchu Grammar*. Brill.

Johannes Heinecke. 2019. ConlluEditor: a fully graphical editor for Universal Dependencies treebank files. In *Universal Dependencies Workshop 2019*, Paris.

Taeho Jang. 2020. Xibe and the manchuric languages. In Martine Robbeets and Alexander Savelyev, editors, *The Oxford Guide to the Transeurasian Languages*, pages 269–287. Oxford University Press.

Gertraude Roth Li. 2000. *Manchu: a Textbook for Reading Documents*. University of Hawaii Press.

Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, et al. 2013. Universal dependency annotation for multilingual parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 92–97.

Christopher Moseley. 2010. *Atlas of the World's Languages in Danger*. UNESCO.

Joakim Nivre, Marie-Catherine De Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, et al. 2016. Universal dependencies v1: a multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666, Portoroz, Slovenia.

Jerry Norman. 2020. *A comprehensive Manchu-English dictionary*. Brill.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2012. A universal part-of-speech tagset. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2089–2096, Istanbul, Turkey, May. European Language Resources Association (ELRA).

Zeeshan Ali Sayyed and Daniel Dakota. 2021. Annotations matter: Leveraging multi-task learning to parse UD and SUD. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3467–3481, Online, August.

Francis Tyers, Mariya Sheyanova, and Jonathan Washington. 2017. UD Annotatrix: an annotation tool for universal dependencies. In *Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories*, pages 10–17.

Ziqing Yang, Zihang Xu, Yiming Cui, Baoxin Wang, Min Lin, Dayong Wu, and Zhigang Chen. 2022. CINO: A Chinese minority pre-trained language model. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3937–3949, Gyeongju, Republic of Korea, October.

Daniel Zeman. 2008. Reusable tagset conversion using tagset drivers. In *Proceedings of the 6th International Conference on Language Resources and Evaluation*.

Veronika Zikmundová. 2013. *Spoken Sibe: Morphology of the Inflected Parts of Speech*. Karolinum Press.

何文勤. 2006. *nimangga gisun* 语文. 新疆教育出版社.

佘吐肯. 2006. 论锡伯文和满文的源流关系. 伊犁师范学院学报: 社会科学版, (4):34–39.

佘吐肯. 2009. *sibe gisun i gisun kooli hafupi leolen* 锡伯语语法通论. 新疆人民出版社.

佟加·庆夫. 1996. 现代锡伯语书面语与口语比较. 满语研究, 22(1):48–60.

国务院第七次全国人口普查领导小组办公室. 2021. 中国人口普查年鉴 *2020*. 中国统计出版社.

安俊. 1985. 锡伯语言文字乃满语满文的继续. 满语研究, 1(1):41–47.

安双成. 1997. 锡伯族与满语文. 满语研究, 25(2):18–26.

张泰镐. 2008. 锡伯语语法研究. 云南民族出版社.

新疆维吾尔自治区民族语言文学工作委员会. 1992. *nei fon sibe šu tacin gisun i arara kooli* 现代锡伯文学语言正字法. 新疆人民出版社.

朝克. 2009. 中国的濒危民族语言文字的保护及学术价值. 中国社会科学院研究生院学报, (4):120–129.

朝克. 2014. 满通古斯语族语言研究史论. 中国社会科学出版社.

胡增益. 2020. 新满汉大词典. 商务印书馆.

苏承志. 1995. 简论现代锡伯语口语和书面语的相异性. 语言与翻译, 43(3):48–56.

赵阿平, 郭孟秀, 唐戈, 吴雪娟, and 杨惠滨. 2003. 满-通古斯语族语言文化抢救调查——新疆察布查尔县锡伯族语言文化调查报告. 满语研究, 37(2):75–81.

顾松洁. 2016. 锡伯语研究综述. 满语研究, 2(2):83–87.

附录

A 论文中涉及的语法注释缩写

缩写	意义	缩写	意义
1	第一人称	IPFV	未完成体
2	第二人称	PFV	完成体
3	第三人称	PROG	进行体
ACC	宾格	NMLZ	名词化标记
GEN	属格	NR.PST	最近过去时
PTCP	形动词	PRS	现在时
COP	系动词	PST	过去时
CVB	副动词	POSS	领属
SG	单数	PTCL	助词
PL	复数	TOP	话题标记

表 A.1: 语法注释缩写

B 树库标注的词性标记及其分布

词性标记	意义	语法例句	新闻	小学课文
NOUN	名词	1344	3262	2363
VERB	动词	1222	1868	1575
PUNCT	标点符号	908	987	1185
ADP	后置词	615	1245	837
PRON	代词	439	190	297
ADV	副词	350	254	381
ADJ	形容词	326	662	556
NUM	数词	145	344	304
PROPN	专有名词	91	431	153
SCONJ	从属连词	80	7	35
DET	限定词	73	33	58
PART	助词	62	71	65
AUX	助动词	55	66	52
X	无类别	33	132	140
CCONJ	并列连词	10	91	60
INTJ	疑问词	4	0	1
SYM	特殊符号	0	1	0
总计	-	5757	9644	8062

表 B.1: 锡伯语树库的词性标记及其分布

C 树库标注的形态特征及其分布

特征	值	语法例句	新闻	小学课文
Abbr	Yes	0	20	0
Aspect	Imp	407	1318	742
	Perf	143	240	371
	Prog	12	9	20
Case	Abl	45	73	48
	Acc	201	556	348
	Dat	226	190	134
	Gen	207	404	330
	Ins	0	12	14
	Lat	13	6	23
	Loc	1	114	139
Clusivity	Ex	9	6	2
	In	1	13	0
Foreign	Yes	0	1	0
Mood	Cnd	67	7	42
	Des	0	0	7
	Imp	103	8	15
	Opt	5	0	15
	Sub	4	1	1
NumType	Card	103	276	269
	Frac	0	2	0
	Mult	8	0	2
	Ord	11	44	20
Number	Plur	95	62	67

表 C.1: 锡伯语树库形态标注及其分布

特征	值	语法例句	新闻	小学课文
	Sing	267	31	153
Person	1	141	38	76
	2	98	4	32
	3	98	26	90
Polarity	Neg	177	36	93
Polite	Elev	1	0	2
Poss	Yes	89	13	68
PronType	Dem	88	36	71
	Int	38	1	29
	Prs	340	69	199
	Tot	11	105	17
Reflex	Yes	14	10	11
Tense	Past	189	168	247
	Pres	368	110	176
Typo	Yes	0	5	0
VerbForm	Conv	414	852	715
	Fin	518	284	444
	Part	177	594	319
	Vnoun	38	127	51
Voice	Cau	40	98	73
	Pass	5	13	13

表 C.1: 锡伯语树库形态标注及其分布

D 树库标注的依存关系及其分布

依存关系	意义	语法例句	新闻	小学课文
acl	形容词性从句	61	232	127
acl:relcl	关系子句	51	205	87
advcl	副词性从句	489	856	775
advmod	副词性修饰语	376	328	429
amod	形容词性修饰语	153	398	313
appos	同位语修饰语	13	78	39
aux	助动词	39	34	23
case	格标记	610	1239	835
cc	并列连词	13	93	60
ccomp	从句补足语	37	26	31
clf	量词	3	63	39
compound	复合关系	188	492	349
conj	并列关系	133	374	187
cop	助动词	9	30	28
csubj	从句性主语	11	0	19
det	限定词	70	121	62
discourse	话语标记	61	8	21
fixed	固定多词表达	4	4	8
flat	平面多词表达	31	152	53
flat:name	人名	2	176	63
flat:num	复杂数字	0	0	55
mark	标记	78	13	34
mark:adv	副词化标记	4	46	35
mark:plur	复数标记	0	5	1

表 D.1: 锡伯语树库依存关系及其分布

第二十三届中国计算语言学大会论文集，第290页-第304页，太原，中国，2024年7月25日至28日。
卷1：主会论文

依存关系	意义	语法例句	新闻	小学课文
mark:rel	形容词化标记	1	4	11
nmod	名词性修饰语	256	1297	537
nmod:poss	具有领属关系的名词性修饰语	89	26	72
nsubj	名词性主语	509	434	603
nsubj:pass	被动结构中的名词性主语	0	7	0
nummod	数词修饰语	99	240	215
obj	宾语	347	754	619
obl	名词性间接格	406	319	547
obl:lmod	方位修饰语	0	103	0
obl:tmod	时间修饰语	41	91	0
parataxis	意合关系	37	81	126
punct	标点符号	908	987	1185
root	根节点	544	266	390
vocative	称呼语	5	0	14
xcomp	控制结构、提升结构等	70	61	62

表 D.1: 锡伯语树库依存关系及其分布