2017年9月 云南师范大学学报 Sept.，2017
第15卷第5期 （对外汉语教学与研究版） Vol. 15 No. 5

第二语言习得研究中语料的基本单位及其在汉语中的切分方法
——以T单位为例*

王亚琼， 冯丽萍
（云南大学 国际学院，云南 昆明 650091；北京师范大学 汉语文化学院，北京 100875）

[摘 要] 基本单位的切分是第二语言语料处理时的一个重要步骤。对于汉语二语习得研究来说，切分基本单位时还需要结合汉语的特点。本文以T单位为例，结合切分标准与切分目的，制定了在汉语的单句、复句、多重复句中切分T单位的方法，重点讨论了紧缩句、流水句及内嵌结构的切分情况。本文的结论是：单句、紧缩句可划为一个T单位，划分复句时应采用“分句判断法”，多重复句可结合分句判断法和层次分析法，流水句需根据分句类型划分，内嵌结构应根据语块化程度划分。

[关键词] T单位；语料切分；汉语
中图分类号： H195 文献标识码： A 文章编号： 1672－1306(2017)05－0001－09

一、引 言

对语料进行切分在早期是会话分析（discourse analysis）中的一个步骤，后来在第二语言习得研究中分析学习者语料时得到采用，切分语料是语料处理过程中广泛使用并且最基础的一步。切分基本分析单位（basic analysis unit）/基本单位（basic unit）（下文统称为“基本单位”）的方法和相关内容在发表的研究成果中并不是最重要的内容，但却是研究过程中一个重要的步骤。从理论上来说，切分基本单位是划定细致观察语料的分析范围；从数据处理的实际操作上来说，是为了下一步进行具体的量化计算做准备。

目前对基本单位的专门研究并不多见，多数实证研究或者直接引用他人对基本单位的定义，或者对此并不做说明；对基本单位进行综合深入讨论的主要有Crookes①、Foster et al.②两项研究，他们主要针对口语语料，全面总结分析了学者们采用的13项基本单位。而对如何在汉语的二语习得研究中划分基本单位这个问题，学界还未给予足够的重视。

本文将以T单位为例，结合汉语的特点明确基本单位在汉语中的划分方法，重点探讨一些特殊结构的划分问题。当需要借助汉语语法研究的成果时，本文采用张斌③主编的《现代汉语描写语法》中的

---

* 作者简介：王亚琼，女，河北张家口人，云南大学讲师，博士，研究方向为第二语言习得。
通讯作者：冯丽萍，女，湖北襄阳人，北京师范大学教授，研究方向为第二语言学习与教学。

基金项目：本研究获得云南大学人文社会科学青年研究基金项目“动态系统理论视角下的汉语中介语系统发展研究”（15YNUHSS002）及国家社科基金“面向第二语言习得的汉语句法复杂度测评指标研究”（14BYY146）资助。

① Crookes, G. The utterance, and other basic units for second language discourse analysis [J]. Applied Linguistics, 1990, 11(2).

② Foster, P., Tonkyn, A. & Wigglesworth, G. Measuring spoken language: A unit for all reasons[J]. Applied Linguistics, 2000, 21(3).

③ 张斌. 现代汉语描写语法[M]. 北京：商务印书馆，2010.

观点作为依据。

二、T单位的切分标准

T单位(Minimal Terminable Unit,T-unit,最小可终止单位)最初由Hunt提出。[①] "T单位是在不留下任何句法不完整的残余片段的前提下,所可能分割到的最小片段。"T单位的定义也有4种不同表达方式,3种是从句法结构构成方面的说明:T单位指主句加上附着或其包孕的从属句或小句下成分;另一种是从切分方式上说明的:切分语料时,在保证不残留片段的前提下,所能切分到的最小单位。Hunt这样解释划分T单位的方法:忽略标点,忽略主句之间的并列连词,将这样的语段切分开,将其切分为最短的片段,只要切分完毕的单位,将首字母大写,末尾加句号或者问号时语法正确,并且没有残留的片段时即可。

Street将T单位定义为"语法上允许加句子语气(标点)的最小单位",这与Hunt的定义是一致的。[②] Foster等人对T单位做了进一步总结:"T单位就是一个主句加上它所有的从属/修饰/联合小句"。[③] 这也说明T单位不包含并列的复合句。

结合以上观点,我们总结出划分T单位时的两种标准。

第一种(记为"T-标-1"),需同时满足以下两个要求:

(1)加句尾语气标点能够形成句法正确的句子的最小片段

(2)除此以外没有残留片段

第二种(记为"T-标-2"),只有一种规定:

主句+从属/修饰/嵌入的小句

在英语中,对于句法正确、成分完整的语言材料,"T-标-1"和"T-标-2"是等价的。"T-标-1"趋向于向外切割语段,只要满足其要求,就尽量多地切分;而"T-标-2"趋向于向内限制,只有满足其要求,才能算作1个T单位。两个标准从两个方向共同规定T单位的边界。无论哪种标准,T单位的最小片段都是1个句子,或者可以通过改变标点成为1个句子的小句,这说明T单位中句法成分要齐全。因此当一段语言材料中有句法成分不全的片段时,必然会被剩余在多个T单位之外。由此可以预见T单位划分对语料,尤其是口语语料的涵盖率不一定能达到100%。这也是T单位提出之后,又出现C单位、AS单位的原因。

Foster等人强调切分基本单位时要依据基本单位所反映出的语言使用者的心理过程,亦即在一个基本单位中,语言使用者所要达成的目标。[④] 因此,基本单位的划定要有相应的心理依据,在心理语言学范畴内的信度和效度也要得到保证,而不是简单地把一段语料切割成若干片段。我们在讨论基本单位的确定方法时,也将充分考虑这一原则。

Endicott从心理语言学角度的解读可以为上述T单位的划分提供支持[⑤]:儿童使用语言是将多个单位的意义转化为言语,要有多轮的心理过程,要多次中断前一轮再重新开始另外一轮,T单位就反映

---

① Foster,P.,Tonkyn,A.& Wigglesworth,G.Measuring spoken language: A unit for all reasons[J].Applied Linguistics,2000,21(3).

② Larsen-Freeman,D.& Strom,V.The construction of a second language acquisition index of development[J].Language Learning,1977,27(2).

③ Foster,P.,Tonkyn,A.& Wigglesworth,G.Measuring spoken language: A unit for all reasons[J].Applied Linguistics,2000,21(3).

④ Foster,P.,Tonkyn,A.& Wigglesworth,G.Measuring spoken language: A unit for all reasons[J].Applied Linguistics,2000,21(3).

⑤ Endicott,A.L.A proposed scale for syntactic complexity[J].Research in the Teaching of English,1973,7(1).

了儿童在一轮心理过程中暂存思想和控制语言的能力。

根据上述分析可知 T 单位具有以下特点：从语言属性来看，T 单位自身句法完整，相应地也保证了语义的相对完整性，并有其语言使用中的心理现实性（psychological reality，详见李行德[①]；袁毓林[②]）；从语料处理的切分操作上来看，T 单位既要尽量切小，又要保证不会过度切分留下句法不完整的碎片（fragment），向内向外两个方向的制约恰好就是确定 T 单位边界位置的参考。

三、汉语中 T 单位的切分方法

根据上一节的分析，要满足“T-标 1”的标准，汉语中 T 单位就是“能够独立成句，也保证剩余部分可以独立成句的最小片段”，此处“能够独立成句”指的是通过将标点改为句号或问号后句法和语义都相对完整。“T-标 2”则对汉语不适用。英语在形态上的丰富信息使得做出这种判断比较简单，而汉语相比之下就有很多需要单独讨论的情况。

根据上文的分析，划分 T 单位的核心在于切分后的片段能否独立成句，并且是否为最小片段。汉语中的单句已经符合以上两条要求，而复句由两个或更多的分句构成，如果每个分句都可以独立成句的话，该复句就不是“最小片段”，需要再分。但如果按照分句切分后，分句无法独立成句，该复句就不能再分。处于单句和复句的模糊边界的紧缩句以及多重复句都需要单独讨论，另外流水句、内嵌结构两种特殊现象也需要单独讨论。

（一）在单句、紧缩句中划分 T 单位

汉语的单句，毫无疑问地划为一个 T 单位；复句则具有划分为多个 T 单位的可能。问题是汉语的单句与复句之间存在模糊的边界（张雪涛、唐爱华[③]）。针对这些“纠结现象”，邢福义[④]主张视具体情况权衡利弊。本研究的关注点在于第二语言学习，因此在出现此类问题时我们会结合汉语第二语言学习与教学的实际来进行论证。对于紧缩句是单句还是复句，或者另外划归一类的讨论还没有一致的结论，三种观点各有依据，各有支持者（详见王姝[⑤]）。

按照语言演化的一般规律来看，紧缩句是一种较为“高级”的语言形式，Givón 总结的句法复杂度变化趋势中有两条规律（单词小句早于多词小句，单独小句早于多重小句）可以证明这一点。[⑥] 因此在划分紧缩句时，要遵循“紧缩句产生的 T 单位——高级语言形式（高级语言水平）——复杂度增加”3 个条件的统一。

从母语习得的规律来看，将紧缩句划为 1 个 T 单位更符合上述 3 个条件统一的原则，体现 T 单位长度增加与语言水平提高的一致趋势。以基本单位的长度来评价语言水平，源于对儿童语言发展的观察：当儿童语言发展到更高级的阶段时，语言中会逐渐出现更多的句法成分，也会出现从句，增加了成分必然会引起句中语言单元数量的增加，表面上体现为基本单位长度的增加。Beers & Nagy[⑦] 认为基本单位长度这个指数反映了学生将各种语法角色（名词、定语、被动式等）全部集中在一个句子中的能力。

---

① 李行德. 语法的心理现实性[J]. 国外语言学，1992，(3).

② 袁毓林. 语言学范畴的心理现实性[J].汉语学习，1993，(4).

③ 张雪涛，唐爱华. 汉语单复句区分问题的理论困惑与解决策略[J].语言教学与研究，2005，(4).

④ 邢福义. 现代汉语复句与单句的对立和纠结[J]. 世界汉语教学，1993，7(1).

⑤ 王姝. 紧缩及其句法语义后果[D].吉林大学，2012.

⑥ Givón，T. Introduction [A]. In Givón，T. & Shibatani，M. (Eds) Syntactic Complexity：Diachrony，Acquisition，Neuro-cognition，Evolution[C]. Amsterdam/Philadelphia：John Benjamins Publishing Company，2009：1～22.

⑦ Beers，S.F. & Nagy，W.E. Writing development in four genres from grades three to seven：Syntactic complexity and genre differentiation[J]. Reading and Writing，2010，24(2).

Rice 等人(Rice et al.[①])的研究为也为这种长度类指标的有效性提供了依据。

从第二语言学习规律来看,紧缩句一般是以"框架结构"或者"固定结构"的形式呈现在教材中的,学习者学习这种紧缩句时并非有意识地先学习一个完整的复句,然后再对其进行紧缩,而是作为一个公式结构来学习,例如:

1)也:你不去我们也去。

2)越……越……:他越说我越生气。

3)一边 A 一边 B:他一边走一边唱。

二语学习者对紧缩句的学习过程更接近单句学习的过程,加之从形式上来说,紧缩句内无停顿标记,更接近单句。因此对紧缩句的划分与单句划分统一起来更符合汉语二语学习的特点。

从汉语二语教学安排来看,刘洁[②]曾经统计过《汉语水平等级标准与语法等级大纲》(以下简称《大纲》)标明的紧缩句和4套常用口语教材中出现的紧缩句的分布情况。33例紧缩句主要集中在丙级和丁级中,可见紧缩句是学习者语言水平到达中高级时的学习内容。同时该研究还对汉语学习者加工紧缩句的实际语言能力做了调查,发现总体上来说,对紧缩句的学习效果和汉语水平趋向一致。因此将紧缩句划为1个T单位更符合紧缩句与中高级语言水平的紧密关系。

由此看来,无论是以语言演化规律为依据,还是以语言学习和教学规律为依据,紧缩句都与较高的语言水平相关,而水平的提高和句法复杂度提高的共现趋势又是主流。以上各方面都支持将紧缩句划为一个T单位,而不必切分开。

从实际测算的角度来看,将紧缩句划为1个T单位,或者划为n个T单位,会带来测算指数数值的变化。最合理的情况是,按照此种划分方式得到T单位之后,与T单位相关的指数数值比按照其他方法划分后获得的数值更大。下面以例4)5)来进行说明:

4)/他既不同意也不反对/

按照4)的划分后,用3个与T单位有关的句法复杂度指数来测算这个紧缩句,获得的数值如下:T单位长度=9、句法成分种类/T单位=3、小句/T单位=2。

5)/他既不同意/也不反对/

按照5)的划分,会得到2个T单位,在依据一段语言样本计算上述3个指数时,要对两个T单位获得的数值求平均数,最终结果为:T单位长度=4.5、句法成分种类/T单位=2.5、小句/T单位=1。

由此可见,以4)的划分方式,会使得紧缩句在测算指数上的得分较高,更加符合紧缩句的复杂度高于最基本句式的事实。

总之,无论是来自理论分析的依据还是来自实际测算的依据,都支持将紧缩句划为1个T单位。

(二)在复句中划分T单位

复句具有划分为多个T单位的可能,但是这并不意味复句一定可以划分开,也不意味着每个分句一定可以划为一个T单位。我们发现,复句中有两个因素制约着分句是否划为独立T单位:第一个因素是分句是否为"独立句段"。胡明扬、劲松[③]将分句分为独立句段和非独立句段两类(本文为求术语的统一,称为"可独立分句"/"非独立分句")。可独立分句的语义、结构都完整,在没有上下文和语境的支

① Rice, M.L., Smolik, F., Perpich, D., Thompson, T., Rytting, N. & Blossom, M. Mean length of utterance levels in 6-month intervals for children 3 to 9 years with and without language impairments[J]. Journal of Speech, Language & Hearing Research, 2010, 53(2).

② 刘洁.《语法等级大纲》中的紧缩复句[D]. 华中科技大学, 2010.

③ 胡明扬, 劲松. 流水句初探[J]. 语言教学与研究, 1989, (4).

持下也可以独立成句。而非独立分句，总是需要一些完句成分（语调、指称、“体”成分、情景说明成分、情态成分、语气成分等）才能保证句法正确。划分后的T单位需具有独立成句的能力，这意味着句法和语义都要能够自足；第二个因素是分句中是否有关联词语。根据关联词语（包括副词、连词、助词、超词结构）的隐现，复句可以分为有标复句和无标复句。“如果复句中分句和分句的关系是确定的，不会有其他解释，就无须使用关联词语，而有些复句如果不用关联词语，分句的关系要么表达不出来，要么意思晦涩，要么有歧义，就必须使用关联词语。”①以上两个因素交叉作用，形成4种类型的分句，见表1：

表1　分句类型与T单位划分

		是否为独立句段	是否带关联词语	划为独立T单位
第一类	分句$_{独立无标}$	+	−	+
第二类	分句$_{非独无标}$	−	−	−
第三类	分句$_{非独有标}$	−	+	−
第四类	分句$_{独立有标}$	+	+	−

第一类，“分句$_{独立无标}$”本身具有独立成句能力，也不受关联词语制约，可以切分为一个独立的T单位。例如：

6)/1 雨停了，/2 天晴了，/3 太阳出来了。/

7)/1 我看《北京青年》了，/2 我只把那一集记住了。/（例句来自本研究收集的语料）

8)/1 我没有别的消息，/2 我没有别的事情。/（例句来自本研究收集的语料）

这一类分句都具有独立成句能力，都可以划为一个T单位。

第二类，“分句$_{非独无标}$”虽然不受关联词语的制约，但是句段本身无法独立成句。例如：

9)/1 他住在华盛顿，2 读寄宿学校。/（例句来自本研究收集的语料）

10)/1 他整天游手好闲，2 也没工作，3 也没对象。/

11)/1 我姐姐叫玛丽，2 今年二十岁，3 在北京上大学。/（例句来自Jiang②）

例9)、10)、11)中每一个分句都不受关联词制约，但是只有第一个分句可以独立成句，后面的分句在独立出现时句法语义都不完整，因此不能在分句之间切分，而是应当将整个复句划为一个T单位。

例11)是Jiang③研究中的例句，Jiang主张将这类复句中的分句分为不同的T单位，因为汉语句子中的主语不是必需的，可以认为句法完整。Jiang的切分方式如下：

12)/1 我姐姐叫玛丽，/2 今年二十岁，/3 在北京上大学。/（3个T单位）

本文对此持不同观点。按照T单位的划分原则，如果不能通过简单加句尾标点就转化为独立的简单句的话，就不能划为一个独立的T单位。9)10)11)都属于“共层”现象（邢福义④），从T单位的心理含义来说，“共层”的结构仍然属于同一话题之内。试比较12)与13)：

13)/1 我姐姐叫玛丽，/2 她今年二十岁，/3 她在北京上大学。/

---

① 张斌.现代汉语描写语法[M].北京：商务印书馆，2010：640～643.

② Jiang, W. Measurements of development in L2 written production: The case of L2 Chinese [J]. Applied Linguistics, 2013, 34(1).

③ Jiang, W. Measurements of development in L2 written production: The case of L2 Chinese [J]. Applied Linguistics, 2013, 34(1).

④ 邢福义. 汉语复句研究[M]. 北京：商务印书馆，2001：563～566.

按照我们在表1中的分类,12)中的2、3都属于第二类"分句非独无标",而13)中的2、3都属于第一类"分句独立无标"。

至于非独立句段不能单独成句的原因,一般都是因句法成分缺失而导致的语义不完整,如12)中的2、3。这一类分句不具有独立成句能力,不应单独划为一个T单位。

第三类,"分句非独有标",首先与第二类"分句独立无标"一样,会由于不具备句法的完整性和语义的自足性而无法成为独立的T单位。此外还会受到关联词语的影响,独立成句的能力进一步降低。这类分句受关联词影响的方式与第四类分句相同。其独立成句的能力最差,不应单独划为一个T单位。

第四类,"分句独立有标"情况要比以上3种复杂一些。

关联词语是复句的语法语义标志,虽然关联词语可能在复句中不出现,我们认为凡是出现了的关联词语就属于复句的一部分,具有和其他句法成分一样重要的地位,并非后加入的插入语。复句的分句之间存在语义的关联性,关联词语的出现就是有目的地强化这种关联。虽然汉语依靠"意合"的方法可以完成很多联结,但是出现了关联词语就意味说话人有意将若干个分句整合到一个复句中,将其作为一个整体语义进行组织和加工,因此关联词语可以视为说话人暂存思想和控制语言的能力较高的信号。

因此,我们认为,与我们在紧缩句部分分析的原理相同,无论是为了提高指标的区分度,还是为了遵循语言学习规律,都不应将这类分句单独划为不同的T单位。

综上所述,以分句的性质和标记作为我们划分汉语复句T单位的标准是一种较为简便易行的方法,同时这种切分原则不会降低区分度,不会违背第二语言学习的规律。本文将这种通过判断分句属性划分基本单位的方法统一称为"分句判断法"。

(三)在多重复句中划分T单位

"分句判断法"同样适用于多重复句的切分,但是多重复句中出现了一个新问题:不能独立成句的分句非独无标、分句非独有标、分句独立有标三类分句不能独立作T单位,需要与临近的分句组合,或依附于临近的分句构成一个T单位。一个多重复句中,除了第一个分句和最后一个分句以外,每个分句都有前后相邻的两个分句,那么归于前还是归于后就成为划分T单位时需要进行的判断。我们认为可以根据分句之间关系的紧密程度来做出判断,例如:

14)/1 冬季日短,2 又是雪天,/3 夜色早已笼罩了全市镇。/

15)1 冬季日短,‖2 又是雪天,|3 夜色早已笼罩了全市镇。

例14)中,分句1和分句3,都属于"分句独立无标",可以独立作为一个T单位,但是分句2是非独立有标分句,不能独立划为一个T单位,需要决定归于前或后,从分句之间的语义关系以及关联词语"又"来看,2与1之间的并列关系要比它与3之间的因果关系更紧密,因此我们认为应将2并入前面的T单位,从2和3之间划分开,成为两个T单位。

可见在处理多重复句时,可以综合运用分句判断法和层次分析法来完成T单位的划分。

(四)在流水句中划分T单位

"流水句"是一类特殊的复句。胡明扬、劲松指出:流水句是"一种在非句终句段也出现句终语调,语义联系比较松散,似断还连的无关联词语复句"①。对于流水句,划分时需主要观察分句的成句能力。胡、劲认为"流水句"中的分句需要是"独立句段",也就是可独立分句。在这个定义下流水句的分句都可以划分为一个T单位。但是吴竟存、梁伯枢则认为流水句的分句可以是不完全的主谓句结构,有很多

① 胡明扬,劲松. 流水句初探[J]. 语言教学与研究,1989,(4).

流水句受主语牵动，承前或者蒙后的某个成分作为分句的主语。[①] 这就形成了“核同质、有核距、有共层”的单复句纠结现象。[②] 本文且不讨论“流水句”对分句的要求究竟应该有多严格，如果以最宽松的标准来看，“流水句”的分句是由表1中的第一类和第二类分句构成的：

16)/1 他可没有透出慌张来，/2 走南闯北的多年了，3 他拿得住劲，4 走得更慢了。/(例句来自老舍《上任》，引自吴竞存、梁伯枢，1999：427)。

例16)中的4个分句分别属于第一类(1、3)和第二类(2、4)。由于第二类的2、4不能单独划为一个T单位，必须归于人相邻的T单位中，4只能并入3所在的T单位中。而2则有两种操作方法。此处需要用到多重复句的划分原则，应用多重复句的层次分析法。从语义上来看，2与3的语义联系稍强于它与1的语义联系，复句的第一层次在1和2之间，因此我们建议从1和2之间划开，把2也归于3所在的T单位中。

本文专门讨论“流水句”的原因有二：一是在少有的几项涉及汉语T单位划分的研究中，曹贤文、邓素娟曾经提出“对于不含任何关联成分的流水句，如一些并列复句和承接复句等，每个分句都可被视为一个T单位”，[③]本文认为在流水句分句资格尚不清晰的前提下，更好的方法还是按照“分句判断法”，用分句的类别作划分T单位的主要依据。

(五)在内嵌结构中划分T单位

在英语中，一个小句(clause)可以成为另一个小句的结构成分，例如关系从句、副词性从句，这就是Hunt定义中“embedded”的情况。嵌入的小句作一个句法成分，本质上还是体词性质的，这种情况在汉语中被称为“内嵌小句”。

在汉语语言学研究的初期，复句理论体系参考印欧语的语言学框架，将内嵌小句视为分句，具有内嵌小句的句子视为复句，但是很快就根据汉语实际进行了修改。现在，内嵌小句做句子成分属于简单句的一个句法成分的观点已经得到普遍认可。

从句法上来看，内嵌小句可以出现在主语、定语、补语、宾语位置，内嵌小句所在的简单句应划分为一个T单位。

内嵌小句作宾语，例如：

17)/苏军攻克柏林，意味着欧洲战场已接近尾声。/(例句来自缪俊，2007)[④]

内嵌小句作定语，例如：

18)/吃过他家米线的人都说好。/

内嵌小句作介词宾语，例如：

19)/每一个初到延安的人都会惊讶于延安市区远不像自己想象的那样落后。/(例句来自缪俊[⑤])

20)/他对于我居然能修好这台闹钟非常惊奇。/(例句来自缪俊[⑥])

内嵌小句作动词宾语，例如：

---

① 吴竞存，梁伯枢. 现代汉语句法结构与分析[M]. 台北：五南图书，1999：431.

② 邢福义. 现代汉语复句与单句的对立和纠结[J]. 世界汉语教学，1993，7(1).

③ 曹贤文，邓素娟. 汉语母语和二语书面表现的对比分析——以小学高年级中国学生和大学高年级越南学生的同题汉语作文为例[J].华文教学与研究，2012，(2).

④ 缪俊. 现代汉语句嵌结构研究[D].华东师范大学博士学位论文，2007.

⑤ 缪俊. 现代汉语句嵌结构研究[D].华东师范大学博士学位论文，2007.

⑥ 缪俊. 现代汉语句嵌结构研究[D].华东师范大学博士学位论文，2007.

21)/雷家乡农民给猪"减肥",比赛哪家养的猪"瘦"。/(例句来自 CCL 语料库)

但是内嵌小句做宾语的句子需要注意以下情况:当主句动词是意向类动词("以为""怕"等)和言说类动词("说"、"讲"等)时,作为宾语的内嵌结构可以大量累积。我们分析语料的目的是考察学习者的语言能力,因此要根据言语能够反映的能力来确定处理语料的方法。

从语言加工的角度来讲,"主语+意向类动词/言说类动词"有成为"语块(chunk)"的趋势,例如"我觉得""他说""我认为"等等。我们认为在划分 T 单位时,应当将"主语+意向类动词/言说类动词"部分视为同词相同的属性。这样处理最大的优点是消除了"主语+意向类动词/言说类动词+大量小句"这样的"虚假高级"结构。例如:

22)/智猪博弈讲的是由两只能根据成本-收益分析做出理性决策的猪,一大一小,共同在一个食槽上吃食。/食槽上有一个按钮,……(例句来自 CCL 语料库)

23)/他说,北京很冷,/白天的气温在摄氏零度以下。/(例句来自缪俊[①])

而对于"主语+意向类动词/言说类动词"整合程度较低的情况,还是应当按照一般句子的划分方法来划分。例如:

24)/他说他也爱她,/这是撒谎。/(例句来自 CCL 语料库)

四、结 语

本文根据汉语的语法属性,结合语言认知与学习的研究成果,在关注指标区分度、可操作性的基础上,提出了 T 单位的划分方法——"分句判断法"。此外本文还讨论了出现在前人研究中的两类结构(流水句、内嵌结构)的切分方法。以上观点总结如表 2:

表 2 汉语 T 单位的切分方法

句法结构		T 单位划分
单句		1T
紧缩句		1T
复句	分句$_{独立无标}$	nT
	分句$_{非独无标}$	1T
	分句$_{非独有标}$	1T
	分句$_{独立有标}$	1T
附: 多重复句——分句判断+层次分析 流水句——根据分句类型 内嵌结构——根据"语块"程度		

我们根据姚双云[②]和张文贤、邱立坤[③]的研究,计算了各类句子在全部句子中的比例,结果为单句占 31.10%,无标复句占 49.40%,含有一个关联词的有标复句占 0.97%,含有多个关联词的有标复句占 18.52%。可见,有 51.60%的句子(单句、多关联词有标复句、单关联词有标复句)要划归一个 T 单位,而 49.40%中的无标复句又会有一部分由于分句是非独立句段而应划归到一个 T 单位中。由此可见,所有

① 缪俊. 现代汉语句嵌结构研究[D].华东师范大学博士学位论文,2007.

② 姚双云. 复句关系标记的搭配研究与相关解释[D]. 华中师范大学博士学位论文, 2006.

③ 张文贤, 邱立坤. 基于语料库的关联词搭配研究[J].世界汉语教学,2007,(4).

句子中，有一半以上的句子属于一个T单位，但是究竟有多少句子需要划分为不同的T单位，可划分为多少个T单位，还需要其他研究（例如对复句中独立语段和非独立语段分布的研究）的数据支持。

The basic unit of the corpus in a SLA study and its segmentation method in Chinese: A study of the T-unit in Chinese

WANG Ya-qiong[1] & FENG Li-ping[2]

(1.School of International Education, Yunnan University, Kunming 650091, China;
2.College of Chinese Language and Culture, Beijing Normal University, Beijing 100875, China)

Abstract: The segmentation of the basic unit is a key procedure in SL corpus processing. The corpus-based segmentation in the SLA study which takes Chinese as the second language should take the properties of Chinese in consideration. In this paper, a method for dealing with T-unit in Chinese is established. The simple sentence, the compound sentence, the multiple compound sentence, the compressed sentence, the run-on sentence and the embedded structure are analyzed. It concludes that as for the simple sentence and the run-on sentence, one sentence should be taken as one T-unit; as for the compound sentence, the properties of clauses are decisive; as for the multiple compound sentence, the analytic hierarchy processing should be additionally employed; and as for the embedded structure, the chucking degree should be taken in consideration.

Key Words: T-unit; corpus-based segmentation; Chinese

[责任编辑：赵昆艳]