

# 韩国语定语从句句法特征分析及其自动识别

安帅飞 毕玉德 张婷

(解放军外国语学院, 河南 洛阳 471003)

**摘要:** 在自然语言处理中, 句法分析研究多集中于单句, 也取得了很大的成功。复句处理仍是 NLP 面临的难点之一, 如何将复句自动离析为单句日益受到研究人员的关注。该文从嵌套类复句入手, 通过分析韩国语定语从句的句法结构特征, 归纳总结其左右边界和内部构成的共现关系规则, 构建定语从句识别规则集, 在语料库中进行匹配运算, 实现了定语从句的自动识别。复句成功离析为单句, 为提高机器翻译等应用系统的效能打下了坚实的基础。

**关键词:** 韩国语; 定语从句; 边界规则; 共现关系; 自动识别

## Syntactic Feature Analysis and Automatic Recognition of Korean

### Attributive Clause

AN Shuaifei, BI Yude, ZHANG Ting

(PLA University of Foreign Languages, Luoyang, Henan 471003, China)

**Abstract :** In the natural language processing, there are more researches based on the automatically syntactic analysis of single sentences, and achieved great success. The processing of complex sentence is still one of the difficulties appeared in NLP. How to automatically divide the complex sentences into single sentences is gradually concerned by the researchers. This paper starts with the embedded complex sentences, by analyzing the syntactic structure of Korean attributive clause, summarizing rules of its left and right borders and the Co-occurrence of internal constitution, building the set of recognition regulation of the attributive clause, performing the matching operation in the corpus, so that reaches the automatic recognition of attributive clause. Complex sentences are successfully parted into single sentences, which can lay a solid foundation for improving the efficiency of machine translation and other application system.

**Key words:** Korean; Attributive Clause; Border Rules; Relation of Co-occurrence; Automatic Recognition

## 1 引言

当前, 语篇层面上的复句处理仍是机器翻译等应用系统面临的难点之一, 如何将复句自动离析为单句成为许多人研究的重点。吴锋文<sup>[1]</sup>回顾了汉语复句二十年前的研究, 概述了邢福义团队的汉语复句信息工程、张仕仁<sup>[2]</sup>在复句“功能结构树”及胡金柱<sup>[3]</sup>在复句关系词提取等的研究工作。韩国语复句处理方面, 刘洋等<sup>[4-5]</sup>利用连接词尾对并列类复句进行“解构化”处理, 提出了对韩汉复句机器翻译的改进建议, 并有效地实现了接续复句的自动提取实验。定语从句属于嵌套类复句, 本文从定语从句入手, 重点分析了如何从嵌套类复句自动离

析出单句的问题。

韩国语句子构成遵循一定的顺序: 单词 (단어) → 语节 (어절) → 短语 (구) → 小句 (절) → 句子 (문장)。句子的构成可以看成是单词搭配单元不断扩大的过程。安帅飞、毕玉德<sup>[6]</sup>根据名词短语的结构特征, 实现了名词短语的自动抽取。本文在此基础上, 将研究单元扩大至小句层面, 在自动离析复句的同时, 也为基于实例的机器翻译、信息检索等应用系统提供了更大的处理单元。

收稿日期: 2017-06-10; 定稿日期: 2017-07-13

基金项目: 国家社会科学基金项目“面向语言信息处理的朝鲜语连接词尾语法知识库研究”(16BYY157)

## 2 韩国语定语从句

韩国语中，仅有一对主谓关系的句子称为单句，有两组或两组以上主谓关系的句子为复句<sup>[7]</sup>。根据语言的递归性，复句又划分为嵌套的包孕句与组合的接续句。韩国语句子分类体系如图 1 所示<sup>[8]</sup>：

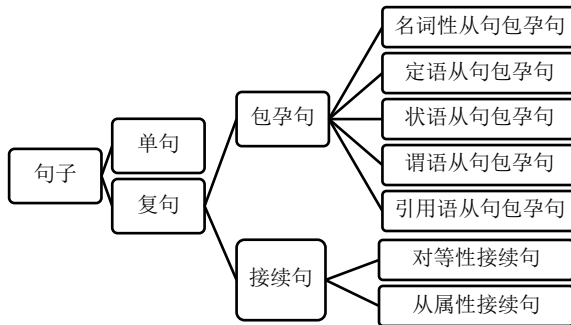


图 1 韩国语句子分类体系图

其中，韩国语包孕句下属的定语从句包孕句即为本文的研究对象<sup>1</sup>。

## 3 韩国语定语从句句法特征及其形式化表示

韩国语定语从句的基本构成为：定语修饰成分、冠形词形词尾、被修饰的中心词。可将其形式化为：AC→AM+ETM+Head<sup>2</sup>。

根据定语修饰成分 AM 与中心词 Head 的关系，可将定语从句分为关系定语从句和同位定语从句<sup>[9]</sup>。

关系定语从句中，中心词 Head 充当定语修饰成分 AM 中的主语、宾语等句子成分。

例 1: 몸에 폭약을 감싼 몇몇 인질들이

나왔다. ( 身上绑着炸药的几名人质出来了。 )

例句中，被修饰的中心词“**몇몇 인질들**(几名人质)”是定语修饰成分 AM 的主语，从句可还原为“**몇몇 인질들이 몸에 폭약을 감싼**

다. ( 几名人质身上绑着炸药) ”。

同位定语从句中，中心词 Head 不作为 AM 的句子成分，与 AM 为同指关系。

例 2: 철수가 합격한 사실을 너는 모르니?

( 你不知道哲洙通过考试了吗? )

例句中，中心词“**사실**(事实)”不是定语修饰成分 AM 的句子成分，而是指 AM 所表示的“**철수가 합격했다.**(哲洙考试及格了)”这一事实。

另外，分析定语修饰成分 AM 的内部构成，可将定语从句分为长定语从句和短定语从句。长定语从句中，定语修饰成分 AM 是整个句子。短定语从句中，定语修饰成分 AM 是主语、谓语、宾语、状语等单句中的句子成分。所有的长定语从句均属于同位定语从句<sup>[10]</sup>。

例 3: 오늘에서야 그가 우리를 위해 애썼

다 (AM) 는 (ETM) 사실을 알았다. ( 直到今天才知道他为我们如此费心。 )

例 4: 예쁜 (AM+ETM) 꽃을 사왔다. ( 买回了漂亮的花。 )

例句 3 中，整个句子作为定语修饰成分 AM；例句 4 中，谓语‘**예쁘다**(漂亮)’作定语修饰成分 AM。

综上，定语从句的分类如图 2 所示。



图 2 定语从句分类图

按照动词中心论观点<sup>[11]</sup>，根据定语修饰成分 AM 中谓词的不同，本文将定语从句分为动词类 AM、形容词类 AM、系词类 AM 定语从句分别进行说明。

<sup>1</sup> 本文仅讨论单句做定语从句的情况，暂不讨论复句作定语(“운동을 좋아하고 공부를 좋아한 철수”)和多重定语(“내 아름다운 신부에게...”)问题。

<sup>2</sup> AC 是定语从句 (Attributive Clause) 的简写；A 是定语 (Attributive) 的简写，M 是 Modifiers 修饰语的简写；ETM 是冠形词形词尾在“韩国语 21 世纪世宗计划”语料标注体系的标注形式。

### 3.1 动词类 AM 定语从句

语料观察实验中, 利用 WordSmith 软件的 Concord 功能, 将关键词设为 ETM, 共现词设为 VV, 从处理结果中选取了 500 句定语从句进行人工观察分析, 归纳总结动词类 AM 定语从句的类型<sup>3</sup>。

#### (1) 关系定语从句

关系定语从句中, 定语修饰成分 AM 最少是一个动词, 冠形词形词尾 ETM 包含: “ㄴ(은) | 던 (过去、回想)、는 (现在)、ㄹ(을) (将来、推测)”。因此其最基本的构成为: VV+ETM+NP; 考虑到在从句末可能会出现 ‘시’、‘았’ 等先语末词尾 EP, 因此其基本构成为: VV (+EP) +ETM+NP。其中, NP 表示单一名词或名词短语。

例 5: 최/NNP 씨/NNB 는/JX 지나/VV ㄴ  
/ETM 5/SN 년/NNB 간/XSN 젊음/NNG 을  
/JKO 바치/VV 았/EP 는데/EC... (崔先生在过去的五年间, 献出了青春...)

除动词之外, 动词类 AM 中往往还含有主语、宾语、状语等。根据语言学规律, 结合在语料库中归纳分析, 关系定语从句的构成可扩展为以下 15 种类型:

#### ① 【主】+VV (+EP) +ETM+NP;

例 6: 이번/NNG 판결/NNG 은/JX 대우  
/NNP 소액/NNG 주주/NNG 들/XSN 이  
/JKS 내/VV ㄴ/ETM 7/SN 건/NNB 의/JKG  
소송/NNG 가운데/NNG..... (这次判决中,  
大宇集团小额股东发起的 7 起诉讼中.....)

主语在语料中的标记形式为:NP+主格助词 JKS。因此, 该类定语从句的形式化表示为:

【NP+JKS】+VV (+EP) +ETM+NP。

<sup>3</sup>形容词类、系词类 AM 定语从句的观察实验与此相同, 下文不再赘述。

#### ② 【宾】+VV (+EP) +ETM+NP;

例 7: 피해/NNG 를/JKO 보/VV ㄴ/ETM  
수재민/NNG 들/XSN 이/JKS 관계/NNG  
당국/NNG 의/JKG 허술하/VA ㄴ/ETM 대  
처/NNG 로/JKB 인하/VV 어/EC 더/MAG  
크/VA ㄴ/ETM 수해/NNG 를/JKO 입/VV  
였/EP 다/EF .SF (因为相关部门的应对不力,  
受灾的灾民遭受了更大的损失。)

宾语在语料中的标记形式为:NP+宾格助词 JKO。因此, 该类定语从句的形式化表示为:

【NP+JKO】+VV (+EP) +ETM+NP。

#### ③ 【状】VV (+EP) +ETM+NP;

例 8: 2/SN 년/NNB 전/NNG 농촌/NNG 에  
/JKB 정착하/VV ㄴ/ETM 남상국/NNP 씨  
/NNB 도/JX 농촌/NNG 생활/NNG 을/JKO  
적응/NNG 하/VV 았/EP 다/EF .SF (2 年前  
在农村安家的南相国(音)适应了农村的生活。)

通过对语料的观察分析, 动词类 AM 中状语主要有 6 种基本表现形式: AVM<sub>1</sub><sup>4</sup>=MAG;

AVM<sub>2</sub>=VV|VA+게; AVM<sub>3</sub>=NP+ (로|에|에게)  
/JKB; AVM<sub>4</sub>=SN+时间类 NNG; AVM<sub>5</sub>=NP  
(+JX) +MAG; AVM<sub>6</sub>=MM+NNG。从句中的状语成分由这六种基本形式单独或组合构成, 将其形式化为[AVM<sub>1</sub>—AVM<sub>6</sub>]。

在定语修饰成分 AM 中, 主语、宾语、状语等会交叉出现, 且韩国语序自由, 各成分位置并不固定。各成分相互交叉, 组合为以下形式:

#### ④ 【主宾】+VV (+EP) +ETM+NP; <sup>5</sup>

<sup>4</sup> AVM 是状语 (Adverbial Modifier) 的简写。

<sup>5</sup> 受篇幅所限, 组合类从句不再举例说明。下同。

在语料中体现为【NP+JKS】+【NP+JKO】+VV(+EP)+ETM+NP。

⑤【主状】+VV(+EP)+ETM+NP;

在语料中体现为【NP+JKS】+【[AVM1—AVM6]】+VV(+EP)+ETM+NP。

⑥【状主】+VV(+EP)+ETM+NP;

在语料中体现为【[AVM1—AVM6]】+【NP+JKS】+VV(+EP)+ETM+NP。

⑦【宾主】+VV(+EP)+ETM+NP;

在语料中体现为【NP+JKO】+【NP+JKS】+VV(+EP)+ETM+NP。

⑧【宾状】+VV(+EP)+ETM+NP;

在语料中体现为【NP+JKO】+【[AVM1—AVM6]】+VV(+EP)+ETM+NP。

⑨【状宾】+VV(+EP)+ETM+NP;

在语料中体现为【[AVM1—AVM6]】+【NP+JKO】+VV(+EP)+ETM+NP。

⑩【主宾状】+VV(+EP)+ETM+NP;

在语料中体现为【NP+JKS】+【NP+JKO】+【[AVM1—AVM6]】+VV(+EP)+ETM+NP。

⑪【主状宾】+VV(+EP)+ETM+NP;

在语料中体现为【NP+JKS】+【[AVM1—AVM6]】+【NP+JKO】+VV(+EP)+ETM+NP。

⑫【宾主状】+VV(+EP)+ETM+NP;

在语料中体现为【NP+JKO】+【NP+JKS】+【[AVM1—AVM6]】+VV(+EP)+ETM+NP。

⑬【宾状主】+VV(+EP)+ETM+NP;

在语料中体现为【NP+JKO】+【[AVM1—AVM6]】+【NP+JKS】+VV(+EP)+ETM+NP。

⑭【状主宾】+VV(+EP)+ETM+NP;

在语料中体现为【[AVM1—AVM6]】+【NP+JKS】+【NP+JKO】+VV(+EP)+ETM+NP。

⑮【状宾主】+VV(+EP)+ETM+NP;

在语料中体现为【[AVM1—AVM6]】+【NP+JKO】+【NP+JKS】+VV(+EP)+ETM+NP。

#### (2) 同位定语从句

同位定语从句分为长定语从句和短定语从句。

a.长定语从句中,定语修饰成分AM中含有终结词尾“다(陈述)、냐(疑问)、자(命令、共动)”三者之一,冠形词形词尾只能是‘는’<sup>[12]</sup>。考虑到在从句末可能会出现‘시’

、‘었’等先语末词尾,因此长定语从句的基

本构成为:VV(+EP)+다|냐|자+는/ETM+NP

。中心词NP与前面的AM为同指关系。通过

对语料的观察分析,该类中心词集中于“사실

(事实)、경우(情况)、일(事)、점(点)

、때(时候)”等。与关系从句相同,其AM

中也含有主语、宾语、状语等,其构成类型同样可扩展15种组合类型,本节不再详述。

b.短定语从句中,定语修饰成分AM中不含终结词尾,中心词Head与长定语从句相同,

基本构成为:VV(+EP)+ETM+NP。短定语

从句的AM、ETM构成与关系定语相同,同样可扩展15种组合类型,不再详述。

### 3.2 形容词类AM定语从句

#### (1) 关系定语从句

关系定语从句中,定语修饰成分AM最少是一个形容词,冠形词形词尾ETM包含:“던

(过去、回想)、ㄴ(은)(现在)”。其最基本的构成为:VA+ETM+NP;考虑到在从句末

可能会出现‘시’、‘었’等先语末词尾。因此其基本构成为:VA(+EP)+ETM+NP。

定语修饰成分AM中,除了基本的形容词之外,往往还含有主语、状语等。因此,关系

定语从句的构成可扩展为以下4种类型:

①【主】+VA(+EP)+ETM+NP。

主语在语料中的标记形式为:NP+主格助词JKS。因此,该类定语从句的形式化表示为:

【NP+JKS】+VA(+EP)+ETM+NP。

②【状】+VA(+EP)+ETM+NP。

通过对选样语料的观察分析,形容词类AM状语也含有6种基本表现形式:

AVM<sub>1</sub>=MAG; AVM<sub>2</sub>=VV|VA+게; AVM<sub>3</sub>=NP+

(로|에|에게|와|과|보다|처럼)/JKB; AVM<sub>4</sub>=SN+

时间类 NNG;  $AVM_5=NP(+JX)+MAG$ ;  
 $AVM_6=MM+NNG$ 。从句中的状语成分由这六种基本形式单独或组合构成, 将其形式化为  $[AVM1-AVM6]$ 。

③【主状】+VA(+EP)+ETM+NP。

该类结构在语料中体现为【NP+JKS】+  
【 $[AVM1-AVM6]$ 】+VA(+EP)+ETM+NP。

④【状主】+VA(+EP)+ETM+NP。

该类结构在语料中体现为【 $[AVM1-AVM6]$ 】+  
【NP+JKS】+VA(+EP)+ETM+NP。

(2) 同位定语从句

同位定语从句分为长定语从句和短定语从句。

a.长定语从句中, 定语修饰成分 AM 中含有终结词尾“다(陈述)、냐(疑问)、자(命令、共动)”三者之一, 冠形词形词尾 ETM 是‘는’。考虑到先语末词尾的出现, 长定语

从句的基本构成为:  $VA(+EP)+다|냐|자+는$

/ETM+NP。中心词 NP 与前面的 AM 为同指关

系。该类中心词也集中于“사실(事实)、경

우(情况)、일(事)、점(点)、때(时候)”等。与关系从句相同, 其 AM 中也含有主语、

宾语、状语等, 其构成类型同样可扩展 4 种组合类型, 本节不再详述。  
b.短定语从句中, 定语修饰成分 AM 中不含终结词尾, 中心词 Head 与长定语从句相同, 基本构成为:  $VA(+EP)+ETM+NP$ 。短定语从句的 AM、ETM 构成与关系定语相同, 同样可扩展 4 种组合类型, 不再详述。

### 3.3 系词类 AM 定语从句

(1) 关系定语从句

关系定语从句中, 系词‘이다’前接名词短语, 构成 AM 中的谓词成分, AM 的基本结构是“NP+O|”。冠形词形词尾 ETM 包含“

던(过去、回想)、ㄴ(现在)”。考虑到先语末词尾的出现, 该类定语从句的基本构成为:  $NP+O|(+EP)+ETM+NP$ 。

定语修饰成分 AM 中, 副词等状语修饰语, 扩展了 AM 的基本结构。关系定语从句的构成便扩展为: 【状】+NP+O|(+EP)+ETM+

NP。经过对选样语料的分析, 发现状语的基本形式主要有 4 种:  $AVM_1=MAG$ ;  $AVM_2=SN+$

时间类 NNG;  $AVM_3=NP+(로|에|에게)/JKB$ ;

$AVM_4=NP+(부터|까지)/JX$ 。从句中的状语成分由这四种基本形式单独或组合构成, 将其形式化为 $[AVM1-AVM4]$ 。

(2) 同位定语从句

同位定语从句分为长定语从句和短定语从句。

a.长定语从句中, 定语修饰成分 AM 中含有终结词尾‘라’, 冠形词形词尾 ETM 为‘는|ㄴ|던’。考虑到先语末词尾的出现, 长定语

从句的基本构成为:  $NP+O|(+EP)+라+는|ㄴ|던/ETM+NP$ 。中心词 NP 与前面的 AM 为同指

关系。该类中心词也集中于“사실(事实)、

경우(情况)、일(事)、점(点)、때(时候)”等。与关系从句相同, 其 AM 中也含有主语、

宾语、状语等, 其构成类型同样可扩展为: 【状】+NP+O|(+EP)+라+는|ㄴ|던/ETM+NP

。

b.短定语从句中, 定语修饰成分 AM 中不含终结词尾, 中心词 Head 与长定语从句相同, 基本构成为:  $NP+O|(+EP)+ETM+NP$ 。短定语从句的 AM、ETM 构成与关系定语相同, 同样可扩展为: 【状】+NP+O|(+EP)+ETM+NP。

#### 4 韩国语定语从句自动识别实验

实验时,按照前述定语从句句法结构特征,归纳分析其在语料中的左右边界规则和内部构成间的共现关系规则,构建定语从句识别规则集。根据识别规则集,对标注语料进行匹配运算,自动识别出定语从句。在此过程中,分析错误的识别结果,迭代完善规则集,最终自动识别出定语从句。具体流程如图3所示。

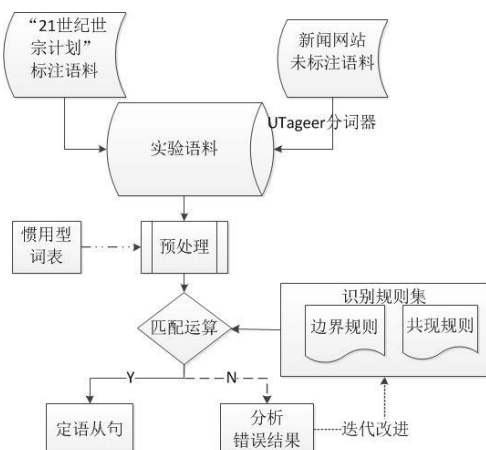


图3 韩国语定语从句自动识别实验流程图

##### 4.1 实验语料及预处理

本文所用语料共80万句,来源于两处:①韩国政府为推动韩文信息化发展,自1998年开始实施、2007年建成的“21世纪世宗计划”标注语料库。该语料库涵盖新闻、小说、杂志等。

本文从中选取了50万句。②网站抓取、后期整理后,获得政治、军事、外交、安全、经济、科技等新闻语句,利用UTageer分词器(标注体系与“21世纪世宗计划”标注语料相同)进行分词处理,得到30万句标注语料。

本文自动识别的对象是定语从句,其基本结构为谓词+ETM+NP。谓词分为单一谓词和复合谓词,在所用的标注语料中,单一动词、形容词被标记为VV、VA,派生动词、派生形容词的标记为NNG--XSV、NNG--XSA,合成动词、合成形容词的标记为VV--EC--VV|VX、VA--EC--VA|VX。为方便后期处理,在实验之初,使用正则表达式将复合动词、复合形容词的标记形式统一替换为VV和VA。

另外,韩国语中存在着“谓词+ETM+NP”类惯用型,标注形式与定语从句相同。实验时,通过匹配惯用型词表将该类惯用型剔除。本文在构建惯用型词表时,有两处来源:①韦旭升<sup>[13]</sup>列举的韩国语常用格式中,将其中的“-르

것-”“-ㄴ 바람”等标记形式为“-ETM+NP”

类惯用型共计40个归入惯用型词表。②在韦旭升的基础上,比照李姬子<sup>[14]</sup>列举的常用惯用型,在词表中补充了25个。文中所用惯用型词表如表1所示。

表1 惯用型词表

序号	-ETM+NP类	序号	-ETM+NP类	序号	-ETM+NP类
1	-ㄴ (-는) 가운데	23	-ㄴ(-은,-는,-르,-을) 바	45	-던 길
2	-ㄴ 감	24	-ㄴ(-은,-는) 바람	46	-는(-은) 물론
3	-ㄴ(-은,-는,-르,-을) 것	25	-ㄴ (-는) 반면	47	-는 한편
4	-ㄴ 게	26	-ㄴ (-는) 법	48	-르(-을) 길
5	-ㄴ(-은,-는) 격	27	-ㄴ(-는,-르) 상	49	-르(-을) 까
6	-ㄴ(-은,-던, -르) 겸	28	-ㄴ(-은,-는) 셈	50	-르(-을) 나위
7	-는 길	29	-는 수	51	-르(-을) 여지
8	-는 김	30	-ㄴ 이래	52	-르(-을) 둥
9	-ㄴ 끝	31	-ㄴ(-은,-는) 이상	53	-르(-을) 따름
10	-ㄴ 나머지	32	-ㄴ (-는) 줄	54	-르(-을) 래
11	-르 나름	33	-는 중	55	-르(-을) 리
12	-ㄴ 다음	34	-ㄴ(-은, -을) 적	56	-르(-을) 마련
13	-ㄴ(-는,-르) 대로	35	-ㄴ 지	57	-르(-을) 만
14	-ㄴ (-는) 대신	36	-ㄴ(-은,-는,-던, -을) 참	58	-르(-을) 뻔
15	-ㄴ(-는,-르,-을) 데	37	-ㄴ 채	59	-르(-을) 뿐

16	-는 도중	38	-ㄴ(-은,-는) 체	60	-ㄴ(-을) 수
17	-는 둘째	39	-ㄴ(-은,-는) 척	61	-ㄴ(-을) 줄
18	-ㄴ(-은,-는) 동시	40	-ㄴ(-은,-는) 탓	62	-ㄴ(-을) 지
19	-ㄴ(-은) 동	41	-ㄴ(-은,-는) 통	63	-ㄴ(-을) 터
20	-ㄴ 뒤	42	-ㄴ (-는) 편	64	-ㄴ(-을) 턱
21	-ㄴ(-는,-ㄴ) 듯	43	-는 한	65	-ㄴ(-을) 말
22	-ㄴ (-는) 마당	44	-ㄴ 후		

## 4.2 韩国语定语从句的识别规则

定语从句的识别规则包含左右边界规则和从句内部结构的共现关系规则。

### 4.2.1 韩国语定语从句的左右边界规则

根据第 3 章分析的定语从句句法结构特征，观察其在语料中的左右边界特征表现，并籍此来界定定语从句。

#### (1) 左边界界定

通过在语料中的观察及实验迭代分析，得出定语从句的左边界存在以下情况：

①句子以定语从句开头，左边紧邻词不存在。

例 9: 신중하/VA ㄴ/ETM 사람/NNG 이  
 /VCP 라면/EC 평범하/VA ㄴ/ETM 이/NP  
 들/XSN 이/JKS 내세우/VV 는/ETM 장점  
 /NNG 을/JKO 감추/VV ㄴ/ETM 것/NNB  
 이/VCP 다/EF /SF (如果说这些人是谨慎的  
 人，平凡的他们有着坚韧的优点。)

②左边界紧邻词为连接词尾 EC

例 10: 건강/NNG 을/JKO 하/VV 기/ETN  
위하/VV 어/EC 부인/NNG 이/JKS 식당  
 /NNG 일/NNG 을/JKO 하/VV 어/EC 벌  
 /VV ㄴ/ETM 돈/NNG 까지/JX 모두  
 /MAG 투자하/VV 였/EP 다/EF /SF (为了  
 健康，连老婆在饭店打工挣的钱都投资了进

去。)

EC 作为连接复句的标志词，可作为其后定语从句的左边界。

③左边界紧邻词为冠形词形词尾 ETM

例 11: 태풍/NNG 루사/NNP 로/JKB 막대

하/VA ㄴ/ETM 피해/NNG 를/JKO 보/VV

ㄴ/ETM 수재민/NNG 들/XSN 이/JKS...

(受台风‘罗莎’影响，而遭受损失的灾民...)

该类定语从句含有双(多)重定语，本文从基本单元入手，分层级解决嵌套问题。

④左边界紧邻词为补助词 JX

例 12: 푸틴/NNP 은/JX 보리스/NNP 열친

/NNP 전/MM 대통령/NNG 이/JKS 손수

/MAG 고르/VA ㄴ/ETM 후계자/NNG 이

/VCP 다/EF /SF (普京是前任总统鲍里斯·叶利钦亲手选定的后继者。)

句中出現两个主语，主句的主语出现在从句的主语前，充当从句的左边界。

⑤左边界紧邻词为主格助词 JKS

例 13: 만약/NNG 북한/NNP 이/JKS 핵

/NNG 생물/NNG 화학/NNG 무기/NNG 와

/JKB 같/VA 은/ETM 대량/NNG 살상/NNG

무기/NNG 를/JKO 사용하/VV ... (万一朝鲜使用了像核生化武器一样的大规模杀伤武器...)

⑥左边界紧邻词为副词格助词 JKB

例 14: ...떠나/VV 기/ETN 전/NNG 에/JKB

필요하/VA ㄴ/ETM 모든/MM 자료/NNG

를/JKO 팩스/NNG 등/NNB 으로/JKB 받

/VV 아/EC ... (在离开前, 得到了所有必需

的资料的传真...)

⑦左边界紧邻词为宾格助词 JKO

例 15: 올해/NNG 를/JKO 복되/VA ㄴ

/ETM 한/MM 해/NNB 로/JKB 기억하

/VV ㄹ/ETM 기업/NNG 도/JX 있/VA 다

/EF /SF (也有企业将今年记忆为幸福的一年。)

⑧左边界紧邻词为逗号 SP、括号 SS、特殊符号 SW 等

例 16: 샤오/NNG 강/JC (/SS 의식주/NNG

가/JKS 해결되/VV ㄴ/ETM 중등/NNG 생

활/NNG )/SS 사회/NNG 와/JC 사회주의

/NNG 현대화/NNG 건설/NNG 을/JKO 위

하/VV 어/EC... (为了小康社会 ( 衣食住行 得以解决的中等生活) 和社会主义现代化建设...)

(2) 右边界界定

①关系定语从句的右边界界定

关系定语从句以名词短语结尾, 因此可将名词短语的右边界作为关系定语从句的右边界。安帅飞、毕玉德将名词短语的右边界归纳为八种: {主格助词 JKS、宾格助词 JKO、副词格助词 JKB、呼格助词 JKV、补格助词 JKC、补助词 JX、肯定指示词 VCP、否定指示词 VC N}<sup>[6]</sup>。本文经过归纳分析, 发现连接助词 (와|

과|랑) /JC 与 ‘중’、‘등’ 依存名词 NNB 在

名词短语中起连接作用, 也充当关系定语从句的右边界。

例 17: 유엔/NNP 이/JKS 강력하/VA ㄴ

/ETM 새/MM 결의안/NNG 을/JKO 내세

우/VV 있/EP 다/EF /SF (联合国制定了强硬的新决议。)

②同位定语从句的右边界界定

同位定语从句中, 경우、때、사실等中心

词不是定语从句的一部分。此时, 中心词前的冠形词形词尾 ETM 可作为右边界。经过在语料中归纳分析, 得到常见的中心词:

{경우、때、사실、목적、점、약점、장점、

가능성、소문、전망、정도、이유}, 并在实

验过程中不断迭代, 补充该中心词集合。

例 18: 한반도/NNP 에/JKB 긴장/NNG 이

/JKS 고조되/VV ㄹ/ETM 경우/NNG 에

/JKB ㄴ/JX... ( 朝鲜半岛局势紧张的情况下...)

#### 4.2.2 韩国语定语从句内部构成的共现关系规则

根据 4.2.1 中的左右边界规则, 得到了基本的定语从句, 但对于含主语、状语、宾语等修饰成分的句子, 无法判断主语等成分归属于主句还是从句。本文辅以定语从句内部构成间的共现关系规则解决这一问题。

(1) 根据语言学特征, 结合在语料中的观察分析, 得到确定的共现关系规则有四条:

①根据左右边界规则抽取出的成分中, 形容词 있다 为定语从句的谓词时, 【NP+JKS】主语、【[AVM1—AVM6]】状语归属于从句。

②根据左右边界规则抽取出的成分中, 如含有两个主语 (出现两个 JKS), 前一个 JKS 标识的主语归属于主句, 后一个 JKS 标识的主语归属于从句。

③根据左右边界规则抽取出的成分，如是同位定语从句，主语、状语、宾语等修饰成分归属于从句。

④根据左右边界规则抽取出的成分中，含有 NP(+와|과)+갈--的定语从句，其前面如出现主语，则主语归属于主句。

(2) 对于无法确定归属的定语从句，计算内部构成成分间的共现频率，根据频率值来近似地估计各成分间的紧密关系，以判断其归属。

以判断【NP+JKS】是否归属于形容词类 AM 定语从句为例进行说明。在形容词类 AM 定语从句中，首先找到主语成分【NP+JKS】，其出现在 ETM 前，将该 NP 赋值为 a1，然后找到定语从句的中心词，将该中心词赋值为

a2，将 AM 中的形容词赋值为 a3。计算并比较共现概率  $\text{Count}(a1, a3) / \text{Count}a1 * \text{Count}a3$  与  $\text{Count}(a2, a3) / \text{Count}a2 * \text{Count}a3$ 。如果  $\text{Count}(a1, a3) / \text{Count}a1 * \text{Count}a3$  的值大于  $\text{Count}(a2, a3) / \text{Count}a2 * \text{Count}a3$ ，则认定主语成分【NP+JKS】与形容词的结合紧密度高于被修饰的中心词，【NP+JKS】归属于定语从句。反之，【NP+JKS】归属于主句。实验时，为解决数据稀疏问题，本文采用了加一平滑，对每个统计项都进行了加一处理<sup>[15]</sup>。

### 4.3 实验结果及评测

根据定语从句的识别规则集，对 80 万实验语料进行匹配运算，实现了定语从句的自动识别。将其中部分结果翻译展示如表 2 所示。

表 2 定语从句自动识别实验结果表

定语从句	译文
핵/NNG 생물/NNG 화학/NNG 무기/NNG 와/JKB 같/VA 은/ETM 대량/NNG 살상/NNG 무기/NNG	像核生化武器一样的大规模杀伤性武器
혐의/NNG 가/JKS 있/VA 는/ETM 사람/NNG	有嫌疑的人
지수/NNG 영향력/NNG 이/JKS 크/VA ㄴ/ETM 대형/NNG 우량주/NNG 의/JKG 주가/NNG	指数影响力很大的大型蓝筹股股价
거동/NNG 이/JKS 불편하/VA ㄴ/ETM 사람/NNG	行动不便的人
성장/NNG 가능성/NNG 이/JKS 높/VA 은/ETM 기업/NNG	增长可能性高的企业
주요/NNG 선전/NNG 활동/NNG 이/JKS 있/VA 을/ETM 때/NNG	有重要宣传活动的时候
내년/NNG 경기/NNG 가/JKS 부진하/VA ㄴ/ETM 경우/NNG	明年经济萧条的情况
이/NP 같/VA 은/ETM 열기/NNG	像这样的热潮
서로/MAG 다르/VA ㄴ/ETM 말/NNG	互不相同的语言
부시/NNP 행정부/NNG 가/JKS 이라크/NNP 공격/NNG 에/JKB 집착하/VV 는/ETM 근본/NNG 이유/NNG	布什政府坚持打击伊拉克的根本理由
삼성전자/NNP 와/JC LG/SL 전자/NNG 가/JKS 요즘/NNG 벌이/VV 는/ETM 논쟁/NNG	三星电子和 LG 电子最近开展的争论
잘/MAG 마련되/VV ㄴ/ETM 기업	准备充分的企业
존중/NNG 받/VV 는/ETM 사회/NNG	得到尊重的社会
우리/NP 위원회/NNG 가/JKS 발표하/VV ㄴ/ETM 내용/NNG	我们委员会发表的内容
얘기/NNG 가/JKS 나오/VV ㄴ/ETM 정도/NNG	传言风起的程度
사안/NNG 이/JKS 생기/VV ㄴ/ETM 때/NNG	案件发生的时候
약속/NNG 을/JKO 어기/VV ㄴ/ETM 기본/NNG 서비스/NNG	违背约定的基本服务
우유/NNG 에/JKB 물/NNG 을/JKO 타/VV 는/ETM 고용인/NNG	往牛奶里掺水的员工
후환/NNG 을/JKO 두려워하/VV ㄴ/ETM 까닭/NNG	害怕后患的缘故
할머니/NNG 가/JKS 손자/NNG 를/JKO 귀여워하/VV ㄴ/ETM 때/NNG	奶奶宠爱孙子的时候
집단/NNG 적/XSN 이/VCP ㄴ/ETM 선택/NNG	集体的选择
독자/NNG 적/XSN 이/VCP ㄴ/ETM 한/MM 분야/NNG	单独的一个领域
신위/NNG 가/JKS 지방/NNG 이/VCP ㄴ/ETM 경우/NNG	神位是纸牌位的情况

좀/MAG 더/MAG 적극/NNG 적/XSN 이/VCP L/ETM 의사/NN	稍微再积极的念头
피해자/NNG 가/JKS 에이즈/NNG 환자/NNG 이/VCP 라는/ETM 사실/NNG	受害人是艾滋病患者的事实

为验证规则的可行性，本文借助了广泛应用于信息检索和统计学分类领域的正确率（P值）、召回率（R值），以及二者的加权平均F值，用来评价实验结果<sup>[16]</sup>。评测时，另外从新闻、小说、杂志3类语料中分别选取了500句进行实验。然后，将人工分析得到的结果与程序自动识别的结果相比较，如表3所示。

表3 实验结果比对表

体裁	自动识别结果		人工计数
	正确数量	错误数量	
新闻	609	66	676
小说	424	42	465
杂志	397	39	439

分别计算P、R、F的值。结果如表4所示。

表4 实验评测结果表

体裁	P值	R值	F值
新闻	90.22%	90.09%	90.15%
小说	90.99%	91.18%	91.08%
杂志	91.06%	90.43%	90.74%
Average	90.76%	90.57%	90.66%

经过比对分析，得到了实验中错误识别的定语从句有以下3个类型：

(1) 特殊符号（SW）导致的错误。

在本文所选用的蔚山大学开发的分析器将%标记为SW。在语料中，SW绝大多数情况下是定语从句的左边界，但也会出现如“올해/NNG 보다/JKB 15/SN %/SW 증가하/VV L/ETM 8/SN 조/NR 8302/SN 억/NR 원/NNB”等情况，此时特殊符号SW为定语从句的一部分，该类定语从句被错误抽取。

(2) ‘상황’等为中心词导致的错误。

自动识别出的“조지/NNP W/SL 부시

/NNP 대통령/NNG 이/JKS 불가능하/VA L/ETM 상황/NNG”从句中，“상황”为中心词结尾的定语从句是关系定语从句，不是同位语从句，“조지/NNP W/SL 부시/NNP 대통령/NNG 이/JKS”本是主句的主语，被错误识别为从句的主语。

(3) 语料标注错误。

例 19: 뉴스/NNG 뒤/NNG 의/JKG 뉴스/NNG 가/JKS 많/VA 다/EC 보/VX 니/EC e/SL 노블리/NNP 안/MAG 슬/VV 는/ETM 입담/NNG 좋/VA 기/ETN 로/JKB 유명하/VA ...

例句中未登录词“안슬/NNP 는/JX”被错

误地标记成“안/MAG 슬/VV 는/ETM”，导致非定语从句被错误抽取。

## 5 总结与展望

本文通过分析定语从句的句法结构特征，对其左右边界和内部构成成分的共现关系进行归纳总结，构建了定语从句识别规则集，实现了定语从句的自动识别。从嵌套类复句中自动离析出定语从句，为提高韩汉机器翻译、信息检索等应用系统的效能打下了坚实的基础。

本文主要讨论了单句作定语从句的情况，针对复句作定语及多重定语问题，以后将做进一步的分析研究。

## 参考文献

- [1] 吴锋文.汉语复句信息处理研究二十年[J].中文信息学报, 2015, 29(1).
- [2] 张仕仁.汉语复句的结构分析[J].中文信息学报, 1994, 8(4): 43-54.
- [3] 胡金柱, 舒江波, 姚双云等.面向中文信息处理的

复句关系词提取算法研究[J].计算机工程与科学.2009, 37(10): 90-93.

E-mail: tinaam@sina.com

- [4] 刘洋, 毕玉德, 李健.基于句法知识的复句解构对韩汉复句机器翻译改进刍议[J].洛阳师范学院学报, 2017, 36(2): 49-53.
- [5] 刘洋, 毕玉德, 李健.基于语言知识的韩国语复句自动识别策略及实现[J].东北亚外语研究, 2017, 17(2): 42-49.
- [6] 安帅飞, 毕玉德.韩国语名词短语结构特征分析及自动提取[J].中文信息学报, 2013, 27(5): 205-210.
- [7] (韩)李翊燮著, 郭一诚等译: 韩国语语法[M].北京: 世界图书出版公司, 2012: 331.
- [8] 교육인적자원부.고등학교 문법[M].서울대학교 국어교육연구소, 2002: 164.
- [9] 张光军, 江波, (韩)李翊燮著.韩国的语言[M].北京: 北京大学出版社, 2009: 311-312.
- [10] 남기심,고영근.표준국어문법론[M].서울:탑출판사, 2008: 381.
- [11] 毕玉德.现代韩国语动词语义组合关系研究[M].北京: 民族出版社, 2005:27-28.
- [12] 고영근,남기심.고교문법자습서[M].서울:탑출판사, 1999: 124.
- [13] 韦旭升, 许东振.新编韩国语实用语法[M].北京: 外语教学与研究出版社, 2006.6: 613-617.
- [14] (韩)李姬子, (韩)李钟禧编著, 张光军等译.韩国语助词和词尾词典[M].北京: 外语教学与研究出版社, 2010.1.
- [15] 宗成庆.统计自然语言处理[M].北京: 清华大学出版社, 2008.5: 78-79.
- [16] 冯志伟, 胡凤国.数理语言学[M].北京: 商务印书馆, 2012: 367.



安帅飞(1991—), 博士研究生, 主要研究领域为自然语言处理、模块识别。  
E-mail: anshuaifei2013@sina.cn



毕玉德(1967—), 教授, 博士生导师, 主要研究领域为自然语言处理、韩国语句法语义学。  
E-mail: biyude@gmail.com



张婷(1984—), 博士研究生, 主要研究领域为自然语言处理、领域本体构建。