Parsing of Korean Based on CFG Using Sentence Pattern Information

Hyeon-Yeong Lee†, Yi-Gyu Hwang††, and Yong-Seok Lee†††

† *Dept. of computer Science, Chonbuk National University, Chonju, 561-756 Korea*
††ETRI Knowledge Mining Research Team, Daejeon , 305-700 Korea
††† Dept. of computer Science, Chonbuk National University, Chonju, 561-756 Korea

Summary
The Korean language has different structural properties than English. English is a more or less fixed word order language, while Korean is a partially free word order language and it controls sentences by limiting the meanings of the predicate. Therefore it is difficult to describe appropriate grammar or syntactic constraint for the Korean. In this paper, CFG-based grammar is described and the way to solve syntactic ambiguity by using syntactic constraint, which was originally sentence patterns information (SPI), is given. SPI is structural patterns of resorted sentence according to the subcategorization of predicate of Korean. In this thesis 39 sentence patterns are used. SPI solve ambiguity of double-object, double-subject or attachment of noun and adverb phrase which appears in the Korean. However the sentence patterns information can't solve every syntactic ambiguity. These sentences are parsed by using semantic markers with semantic constraint. Semantic markers can be used to solve ambiguity caused by auxiliary particle or commutative case particle. By empirical results of parsing 1000 sentences, we found that our method decreases 88.32% of syntactic ambiguities compared to the method that doesn't use SPI and split the sentence with basic clauses.

Key words:
Resolution of Syntactic Ambiguity, Unification based CFG, Sentence Patterns Information (SPI), Semantic Marker, Parsing

1. Introduction

In Korean, predicate dominates the sentence by constraining the noun phrase with semantics. Particles and endings, which play a functional role in Korean, are fluently cultivated and most of the sentences have relative clauses. These phenomena cause a phrase attachment problem in the syntactic analysis. Therefore, Korean is not like western languages, which have precise grammar rules. Korean is analyzed by the strict constraint, which is the knowledge of the context sensitive meanings. In this point, the grammar rules should be described in a simple way and the way to check and analyze the relation of each morpheme on the process of syntactic analysis is desirable.

However, the most of previous Korean parsing method was used to analyze Korean by using the parsing framework of western languages. Unification based context free grammar(CFG) theories[1,2] are the ways to pick ungrammatical sentences using any conditions of constraint. These theories, however, were difficult for analysis of Korean which has partially free word-order and it's meaning is important. Also, dependency grammar (DG)[3] was developed to resolve ellipses and free word-order which are characteristics of Korean. But, parsing with DG causes over-generation of parse trees which can be avoided by simple phrase structure rule. For this reason, there hasn't been a standard of parsing of the Korean so far. Therefore, we describe the way to identify and resolve the causes of syntactic ambiguity, which appears in parsing of Korean.

The most of syntactic ambiguity appears according to the attachment of predicate and noun phrases, “NP (Noun Phrase) + VP(Verb Phrase)” or “VP + NP”. For example, the noun phrase ‘학교에(hak-kyo-e: to school)’ can be attached to both predicate ‘가는(ka-nun: go)’ and ‘보았다(po-ass-ta: see)’ in <Figure 1>. But, we can easily find that it will be attached to the predicate ‘가는(ka-nun)’ by the semantic meanings of ‘가는(ka-nun)’ and ‘보았다(po-ass-ta)’. But, if we classify the predicate by usage of structural type in the sentence, we can disambiguate this attachment problem in the phase of parsing.

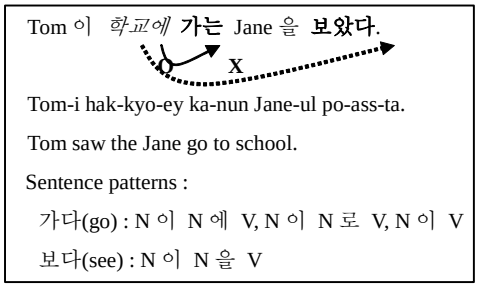


Fig. 1 Example of Attachment Problem

Sentence patterns information (SPI) is called structural type of sentence. In this paper, attachment problem of syntactic ambiguity is solved by using SPI,

Manuscript received July 5, 2007
Manuscript revised July 25, 2007

which is classified for characteristics of Korean from subcategorization information of the predicates. In addition, there are many sentences which have a syntactic ambiguity and this can not be solved by the SPI only. In such case, semantic markers(SM) which have meaning constraint for predicate will be the only possible alternative.

In the Korean parsing, the reason of syntactic ambiguity can be largely classified into two categories. One is morphological ambiguity and the other is caused by attachment problems. Morphological ambiguity, which comes from the result of morphological analysis, can be solved by syntactic morpheme, which is suggested by [4]. But attachment problem caused by the syntactic characteristics of Korean is difficult to solve. Therefore, we describe the syntactic characteristics of Korean in the point of parsing. And, we propose an unification based parsing method using sentence patterns to solve the syntactic ambiguity of Korean.

2. Property of Korean: In the point of syntactic analysis

2.1 Morphological Property of Korean

Functional morpheme has fluently cultivated in Korean and some morphemes often combine to make a syntactic unit. These morphemes are the reasons of morphological ambiguity and syntactic ambiguity. Therefore, many researches [4,5,6] have been done to solve them. [4] suggest syntactic morpheme which is the combination of associated functional morphemes. According to this study, syntactic morpheme can improve the efficiency of syntactic analysis because it can be a syntactic unit for parsing.

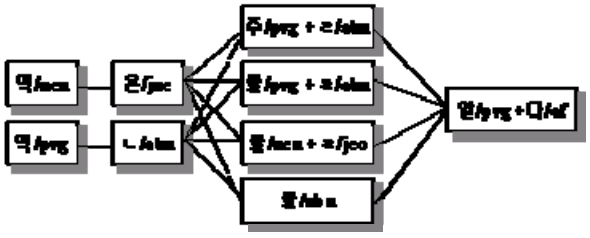


Fig. 2 Result of morphological analysis for “먹은 줄 알다”

The result of the morphological analysis, Fig. 2 above, is for "먹은 줄 알다([I guess] you eat)". This Fig. 2 has 8 morphological ambiguities. If we use syntactic morphemes suggested by [4], modality 'Guess' is described by a combination of morphemes "ㄴ 줄 알다(guess)". Therefore, the only result "먹다(pvg[Guess])" can be obtained. Like above, these syntactic morphemes help to solve syntactic ambiguity. So, syntactic morphemes are used as input data of syntactic analysis in this paper.

2.2 Syntactic property of the Korean

Korean is a non-structured language, which has ellipses and free word order partially and needs a lot of case particles and noun phrases for the predicate. So, it is impossible to use the fixed type of syntactic information only to identify the structure of a sentence. For example,

1) 탐이 귀찮게 군다.
   Tom-i kwi-chan-key kun-ta.
   Tom behaves annoyingly.
2) 탐이 군다.*
   Tom-i kun-ta.*
   Tom annoys[?].*

at above sentence, 1) and 2) "군다(kun-ta: annoy)" is an intransitive verb so a subject can be the essential element. Therefore, 1) and 2) are analyzed to be correct. But the predicate "군다" needs an adverb for "어떠하게 (e-tte-ha-key: how)" as an essential element. So, 2) is not a correct sentence. This situation is not limited to the predicate "군다". There are many predicates which need adverbs and adverbial case particle.

Thus, there are many predicates, which need adverbs and special case particle. Other optional cases are understood as an auxiliary meaning of the Korean. It causes a difficulty of identifying the meaning of a sentence and it may give rise to ambiguity. Therefore, it is necessary to constrain the syntactic type of the predicate. This is called SPI[7]. It is considered that the use of the SPI in syntactic analysis is essential.

Also, there are many sentences, which have two more than predicate. In these sentences, noun phrases and adverbial can be attached to all possible predicate. It is called an attachment problem and this causes syntactic ambiguities mainly in Korean parsing. For example, Fig. 1 shows this NP attachment that the noun phrase ‘학교에(hak-kyo-ey: to school)’ can be attached to both predicate ‘가다(ka-ta: go)’ and ‘보다(po-ta: see)’. But, in the relative phrases of Korean sentence, NP followed predicate play an important role in essential case of the predicates.

Therefore, Fig. 1 can be analyzed for meaning of “Jane이 학교에 가다(Jane-i hak-kyo-ey ka-ta: Jane go es to school)” and then “Tom이 그 Jane을 보다(Tom -i ku Jane-ul po-ta: Tom see the Jane)”. Also, we can know that it will be attached to the predicate ‘가다(k a-ta)’ by the sentence patterns information of ‘가다(ka

-ta)' and '보다(po-ta)'. And then, because "Jane 이 학교에 가다(Jane-i hak-kyo-ey ka-ta)" is satisfied to SPI of '가다(ka-ta)' with "N 이 N 에 V", the NP 'Tom 이(Tom-i)' can avoid to attach to predicate '가다(ka-ta)'. Thus, if we constraint that predicate must satisfy maximum predicate-argument by using the sentence patterns information, we can disambiguate this attachment problem in the phase of parsing.

However, there are situations that it is difficult to solve syntactic ambiguity with the SPI only. For example, in the Fig. 3, if the SPI be used, '아동작가로(a-tong-cak-ka-lo: juvenile novels writer)' can combine with '유명한(yu-myeng-han: famous)' or '철수하였다(chel-su-ha-yess-ta: withdrawn)'. So, this sentence has a syntactic ambiguity.

(a-tong-cak-ka-lo yu-myeng-han cang-kun-i kun-tai-lul chel-su-ha-yess-ta.)

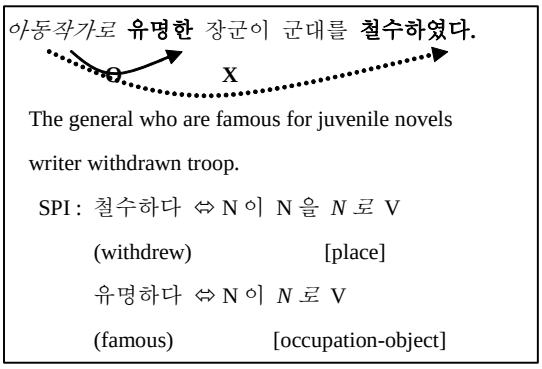


Fig. 3 Examples of SPI and SM

But, it can be solved if a noun phrase, which is in a sentence, is constrained by meaning. In the SPI of predicate "유명하다(yu-myeng-ha-ta: famous)", the semantic type of "N 로(for N)" must be a 'occupation-identity'. So, '아동작가로' must be combined with predicate '유명한' not with predicate '철수하였다'. Semantic marker(SM) is the information which constraint noun phrases in SPI. Syntactic ambiguity is solved by using SM in the case, which it is impossible to solve by the SPI only in this paper. A lot of syntactic ambiguity can be solved with the SPI and SM as shown above.

3. Sentence Patterns in CFG

3.1 The information of sentence patterns(SPI)

A SPI means a sentence template of an essential element to the commonality of a structural type of a sentences[7]. The Korean has a predicate-centered sentence structure which means the sentence structure is identified by predicates not noun phrases. Therefore sentence patterns are classified by predicates. 31 verbs SPI and 8 adjectives SPI are used in this paper. SPI, which are classified, are shown as below.

Table 1: Classified Sentence Patterns Information

V1)	N(이/는/은/가) + V
V2)	N(이) + N(에/에게) + V
V3)	N(이) + N(로/으로) + V
V4)	N(이) + N(와/과) + V
	:
A5)	N1(이) + N2(이) + A
A6)	N(이) + N(로) + A
A7)	N1(이) + N(로) + N2(이) + A
A8)	N1(이) + N(와) + N2(이) + A

3.2 Classification of sentence patterns

The Korean sentence is consisted of complements and modifiers. The complement is essential to make a sentence but the modifier is not essential. The principles below are used to distinguish complements from modifiers to decide which sentence pattern a sentence has.

Principle 1) ***Satisfaction of syntactic/semantic requirements in the predicate****:*

- The complement should satisfy syntactic and semantic requirements of predicates.

For example, in the sentence "Tom 이 Jane 과 싸웠다(Tom-i Jane-kwa ssa-wess-ta: Tom fought with Jane)" the predicate '싸우다(ssa-wu-ta: fight)' needs 'N 와(wa: with N)' for its complement.

- Tom 이 *Jane 과* 싸웠다.
  Tom-i *Jane-kwa* ssa-wess-ta.
  Tom fought *with Jane*.

- Tom 이 싸웠다.*
  Tom-i ssa-wess-ta.*
  Tom fought.*

Principle 2) ***Improperness of ellipses***:

- Complements can not be omitted.

- Tom 이 *성가시게* 군다.
  Tom-i *seng-ka-si-key* kun-ta.
  Tom behaves *annoyingly*.

If the adverbial phrase '성가시게(seng-ka-si-key: annoyingly)' is omitted, then this sentence is ungrammatical. So, the phrase '성가시게' is complements.

Principle 3) ***Improperness of repetition***:

- *A complement, which is used as a special case, can not be used twice in a sentence.* Exceptionally, dual-subject and dual-object, which can be used twice in a sentence, are allowed and it can be solved by SPI. A predicate '되다(toy-ta: become)' has a SPI "N 이 N 이 V".

- *Tom 이 선생님이* 되었다.
  Tom-i sen-sayng-nim-i toy-ess-ta.
  Tom became a teacher.

Principle 4) ***Improperness of inversion***:

- *When a word order is inversed and the sentence does not make sense, this word is a complement. In the following examples, the first sentence is correct in the point of literary style.*

Tom-i *Jane-ul mye-nu-li-lo* sam-ass-ta.
Tom makes *Jane his daughter-in-law.*
- Tom 이 *Jane 을 며느리로* 삼았다.

- Tom 이 *며느리로 Jane 을* 삼았다.*
Tom-i *mye-nu-li-lo Jane-ul* sam-ass-ta.*
Tom makes *his daughter-in-law Jane*.*

So far, it is explained how can we classifies predicates. However there are some problems in analyzing sentences in the Korean with SPI only. Although some predicates have a similar semantic attribute, these predicates may have different SPI in Korean. The constraint for nouns is different even in the same sentence patterns. So, constraint of nouns should be considered with sentence patterns.

For example, verbs of perception - 맡다(math-ta: smell), 시청하다(si-cheng-ha-ta: watch), 보다(po-ta: look) - have the sentence structure "N 이(subject) N 을(object) V". However, nouns for the object have constraints according to the predicate. predicate '시청하다(si-cheng-ha-ta)' and '보다(po-ta)' need '구체물(ku-chey-mul: a specific thing)' but predicate '맡다(math-ta)' needs '추상물(chu-sang-mul: an abstract thing)' or '냄새(naym-say: scent)'. Semantic markers for these nouns are necessary to limit sentence patterns.

맡다 : Tom 이 *냄새를* 맡다.
math-ta : Tom-i *naym-say-lul* math-ta.
Smell : Tom smells *smell*.

시청하다 : Tom 이 *TV 를* 시청하다.
si-cheng-ha-ta : Tom-i *TV-lul* si-cheng-ha-ta
Watch : Tom watches *TV*.

보다 : Tom 이 *신문을* 보다.
po-ta : Tom-i *sin-mun-lul* po-ta
Look : Tom looks at *a newspaper*.

SM is mostly showed with co-occurrence information. However, the co-occurrence information from corpus might cause a data sparseness problem. This means only partial co-occurrence of adverbs, nouns, and predicates. The SPI and SM, which were classified in this paper, can solve the problem of data sparseness more or less. The SM of nouns-predicates and adverbs-predicates is constructed by referring the part of [8].

3.3 Context Free Grammar with Conditional Unification

Conditional unification based CFG is used as a basic framework for syntactic analysis. We describe grammar rules in a simple phrase structure and use conditional unification with SPI and SM to check the relation of each phrase. The examples below show the necessary constraint using the information of sentence patterns and semantic knowledge to apply a phrase structure, "VNP <-> NP VNP".

Table 2: Examples of grammar using SPI and SM

```
(<VNP> <==> (<NP> <VNP>)    ;;; CFG rule
 ((x0 = x2)
  (*or*
   (((x1 jform) =c jcs)
    (*or*
     (((x0 topic) =c subj)
      ((x0 sp-info) =c v6 )          ;; SPI constraint
      ((x0 subj) = *undefined*)
      ((x0 comp) = *undefined*)
      (*or*
       (((x1 sm-info) =c ANI)        ;; SM constraint
        ((x0 subj) = x1))
                :
```

CFG based grammar is characterized by PATRII and this is translated to the GLR parsing table and conditional constraint function for syntactic analysis [9].

4. Parsing a Sentence with the SPI

4.1 Resolution of ambiguity with SPI

In English, the most ambiguous part of the syntactic analysis is prepositional phrase(PP) attachment and coordinate conjunction. Similar to English, adverbial phrase attachment and commutative case particle attachment is very often in Korean. Sentence patterns can solve the problem of adverbial phrase attachment and the

ambiguity caused by the commutative case particle, '와(with)'.

For example, ambiguity with both adverbial phrase attachment and commutative case particle attachment a ppeared at Fig. 4. The commutative case phrase 'Sam 과(kwa: with)' can modify '싸우다(ssa-wo-ta: fight)' o r '보다(po-ta: see)', but the predicate '싸우다(ssa-wo-ta)' has the sentence pattern "N 이(i: subject) N 과(kw a: with) V" and '보다(po-ta: see)' has "N 이(i: subjec t) N을(ul: object) V". So, the phrase 'Sam과(kwa: wi th)' must be combined with '싸우다(ssa-wo-ta)'. Also, we can resolve adverbial phrase attachment problem b y allowing relative clause can have a maximum essen tial argument using SPI. Therefore, because adverbial phrase '학교에서(hak-kyo-ey-se: in the school)' is exi sted between 'Sam 과(kwa)' and '싸우다', it must be attached to '싸우다'.

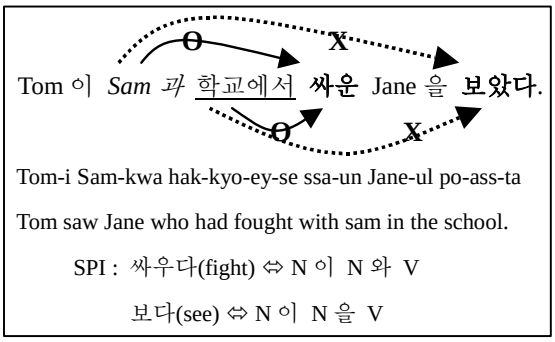


Fig. 4 Examples of syntactic ambiguity

Also, Korean has a commutative case particle attachment problem. Because particle '와/과(wa/kwa: with)' can be regarded conjunctive particle or commutative case particle. A noun with conjunctive particle will be combined another noun. A noun with commutative case particle will be combined a predicate. So, the discrimination of particle '와/과(wa/kwa: with/and)' is only be determined by predicate. This particle '와/과' is a essential element of the SPI. The example is as follows.

Tom 이 *Jane 과* 싸웠다. [N 이 N 와 V]
Tom-i Jane-kwa ssa-wess-ta.
Tom fought with Jane.

Tom 이 *Jane 과 빵을* 먹었다. [N 이 N 을 V]
Tom-i Jane-kwa ppang-ul mek-ess-ta.
Tom and Jane ate bread.

Also, dual-subject of dual-object makes more difficulty in parsing Korean. But, this problem can be resolved by SPI. For example,

Tom 이 돈이 모자랐다. SPI⇔ N 이 N 이 V
Tom-i ton-i mo-ca-la-ta.
Tom is to be not enough money.

Functional words such as particles and endings are richly cultivated in Korean. However, it is not easy to determine the role of an auxiliary particle. Also the commutative case particle '와/과(and, with)' can cause ambiguity depending on how it is combined. These problems can not be resolved by SPI only. So, SPI cooperate with a SM for decreasing ambiguity. For Example, in the next sentence, phrase 'Jane 과(with Jane)' does not combine with '빵(bread)' because of different SM. So, 'Jane 과' plays the role of a subject by combining with 'Tom 이'.

Tom 이 *Jane 과* 빵을 먹었다.
Tom-i *Jane-kwa* ppang-ul mek-ess-ta.
Tom *and Jane* ate bread.

SM : [Tom:Human], [Jane:Human], [Bread:Food]

4.2 Experimental Results

Complex sentences, which are consists of more than 10 words, were chosen for the experiments. The 700 sentences for the test were chosen from KAIST corpus and 300 sentences from a social textbook for an elementary school. Conditional unification based CFG was used as a grammar rule to do parsing in the Korean. Sentence patterns information and semantic marks were used as conditional constraints. The results are shown below.

Test1 : Not use SPI and SM
Test2 : Use SPI
Test3 : Use SPI and SM

Table 3 : Experimental results

Test Set	Average of Predicate Num.	Test1	Test2	Test3
KIBS(700)	3.55	68.43	18.21	6.93
S.T(300)	2.47	51.25	21.04	7.04
Average	3.01	59.84	19.63	6.99

In the results, when SPI and SM are used, the average of ambiguity number is decreased a lots. This means that sentence patterns can be used to resolve the attachment problem in Korean sentence.

5. Conclusion

Conditional unification based CFG is used to do parsing in the Korean. SPI is used to identify Korean, which has partially free word order. Sentence patterns are good constraint for parsing in dealing with attachment problem in relative sentence. Also, SPI are good information for processing in adverbial phrases, a commutative particle, dual-subject and dual-object. By empirical results of parsing 1000 sentences, we found that our method decreases 88.32% of syntactic ambiguities compared to the method that doesn't use SPI and split the sentence with basic clauses.

Conditional unification based CFG using sentence patterns as a constraint can do parsing in the Korean with efficiency. This means a language like Japanese, which is difficult to describe the grammar, can do efficient parsing as long as sentence patterns are identified.

Our future works includes improving SPI for capturing several phenomena of Korean and constructing SM more detail for constraining SPI precisely.

References

[1] S. W. Yang, A Syntactic Analysis for Korean using PATRII based on Conditional Unification, Doctoral dissertation, Chonbuk University, 1995(in Korean).

[2] Tomabechi, H., Efficient Unification for Natural Language, Doctoral dissertation, Carnegie Mellon University, 1993.

[3] Y. D. Yoon, Y. T. Kim, "The Korean Language analysis algorithm based on the dependency grammar using the multi-phase filtering and searching method," Journal of the Korea Information Science Society, Vol 19, No. 6, pp. 614-624, 1992(in korean).

[4] Y. G. Hwang H. Y. Lee and Y. S. Lee 1999 *Resolution Strategy of Morphological Ambiguity for Korean Parsing*, Proceedings of the International Conference on Computer Processing of Oriental Languages, pp. 53-58,

[5] S. S. Kang 1993 *Korean Morphological Analysis Using Syllable Information and Multi-word Unit Information*, Doctoral dissertation, Seoul University (in Korean).

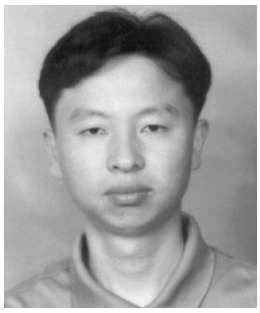
[6] C. J. Kim, C. y. Jung, Y. H. Kim and Y. H. Seo, 1995 *An Efficient Korean Syntactic Analyzer Using Partial Combination of Words*, Proceedings of The 22nd KISS Fall Conference 22/2, pp. 597-600 (in Korean).

[7] U. K. Kang 1995 *A Study on Korean Sentence Patterns* Parkijeong Pub. Com., 1995(in Korean).

[8] Youn-Sai Univ. 1998 *YOUNSAI KOREAN DICTIONARY* Doosan Dong-a Pub. Com. (in Korean).

[9] S. W. Yang 1995 *A Syntactic Analysis for Korean using PATRII based on Conditional Unification*, Doctoral dissertation, Chonbuk University (in Korean).

HyeonYeong Lee received the B.E. and M.E. degrees in Computer Science and Statistics from Chonbuk National Univ. in 1991 and 1996, respectively. He is currently pursing his Ph.D. degree in Chonbuk National Univ. his research interests are Korean Language Processing, and Information Retrieval.



YiGyu Hwang received the BS degree in computer science from Chonbuk National University, Korea, in 1993. He received the MS and PhD degrees in computer science and statistics from Chonbuk National University, Korea, in 1995 and 2001, respectively. Since 2001, he has been with Electronics and Telecommunications Research Institute (ETRI), Korea, as a senior member of research staff. His research interests are natural language processing, information retrieval, and text mining.

YongSeok Lee received the B.E. degree in Electronic Engineering from Seoul Univ. in 1977, M.S. degree in Computer Science from KAIST, Ph.D. degree in Intelligent and Information System Engineering from Tokusima University of Japan in 1995 respectively. From 1979 to 1983, he was a senior researcher at Korea Research Institute of Standards. Since 1983, he has been a professor in the Dept. of Computer Science, Chonbuk National University. His research interests are Korean Language Processing, and Information Retrieval.