A Local Grammar-based Approach to Recognizing of Proper Names in Korean Texts

Jee-Sun NAM* **, Key-Sun CHOI*

nam@world.kaist.ac.kr, kschoi@world.kaist.ac.kr
* CAIR, Department of Computer Science, KAIST, Korea
** IGM, University of Marne-la-Vallee, France

Abstract

We present an LG-based approach to recognizing of Proper Names in Korean texts. Local grammars (*LGs*) are constructed by examining specific syntactic contexts of lexical elements, given that the general syntactic rules, independent from lexical items, cannot provide accurate analyses. The *LGs* will be represented under the form of Finite State Automata (FSA) in our system.

So far as we do not have a dictionary which would provide all proper names, we need auxiliary tools to analyze them. We will examine contexts where strings containing proper names occur. Our approach consists in building an electronic lexicon of *PNs* in a way more satisfactory than other existing methods, such as their recognition in texts by means of statistical approaches or by rule-based methods.

1. Introduction

In this paper, we present a description of the typology of nominal phrases containing Proper Names (PN) and the local grammars [Gro87],[Moh94] constructed on the basis of this description. The goal is to implement a system which detects automatically *PNs* in a given text, allowing the construction of an electronic lexicon of *PNs*.

The definition of Proper Nouns, as opposed to that of Common Nouns, is often a problematic issue in linguistic descriptions [Gar91]. *PNs* are understood in general as phonic sequences associated with one referent, without any intrinsic meanings, such as *Socrates, Bach* or *Paris*. They usually are characterized by nominal determination, the upper case marker, prohibition of pluralizing procedure, or non-translativity [Wil91]. However, semantic or syntactic criteria do not allow to distinguish these two categories in an operational way. For example, nouns such as *sun, earth* or *moon*, semantically appropriate to the definition of proper nouns such as *Mars* or *Jupiter*, do not have to be written with the upper case initial: hence, they are not considered as proper nouns. On the contrary, some proper nouns such as *Baudelaire* or *Napoleon* can be used as well as common nouns in contexts where they occur in metonymic or metaphorical relations with common nouns like:

I read some (Baudelaire + poems of Baudelaire)
He is a real (Napoleon + general)

Moreover, they often allow, like common nouns, the derivation of adjectives: *e.g. Socratic, Napoleonic* or *Parisian*. These are also written with initial upper case, differently from usual adjectives.

The situation concerning French is similar to that. Let us consider [Gar91]:

J'ai écouté (du Bach + de la musique)
J'ai bu (du Champagne + du vin rouge)

Derivational productivity is also underlined: *socratique, parisien* or *newyorkais,* which however do not begin in the upper case.

In the case of English or French, one could delimit formally the category **proper nouns** by means of the upper case, even though this criterion does not correspond entirely to our intuition about proper nouns. However, in Korean, there are no typographical markers such as upper case *vs.* lower case, while one assumes that the nouns such as in (1) could be semantically and syntactically different from those of (2):

(1) 김 민우, 서울. 프랑스
Kim Minu, Seoul, France

(2) 남자, 수도, 나라
namja[man], *sudo*[capital], *nala*[country]

This situation makes more difficult the distinction between proper nouns and common nouns than in the case of French or English, when the former appears in the same grammatical positions as the latter like:

나는 (모짜르트 + 클래식음악)-을 하루종일 들었다
(I listened to (Mazart + classic misic) all day)

그는 (보르도 + 적포도주)-만 마신다
(He only drinks (Bordeaux + red wine))

The derivation of some other categories from *PNs* is also observed:

박정희 - 박정희식	[in Park JungHee's manner]
프랑스 - 프랑스풍	[France style]
중국 - 중국어	[Chinese (language)]

In fact, the distinction between these two categories might be arbitrary. We should perhaps consider a continuum of the noun system: a thesaurus of nouns constituted of the most generic nouns to the most specific nouns (which we call proper nouns). The following example shows a part of a noun hierarchy (Figure 1):

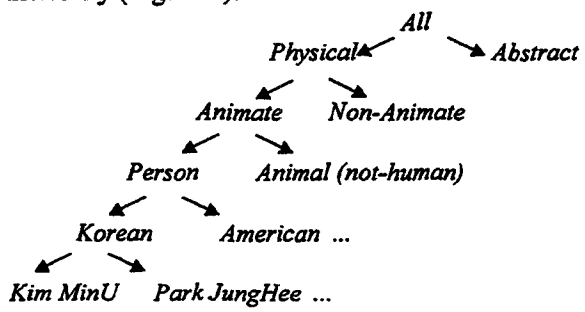
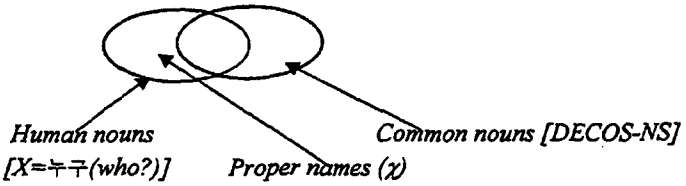


Figure 1

Therefore, in the automatic analyses of texts written in Korean, we intend to consider the definition problem of proper nouns from a different view point: whatever the given definition of proper nouns is, once a complete list of them is available, we presumably do not need any longer this particular distinction between proper and common nouns. All nouns have some semantic and syntactic properties, which lead to group them into several classes, not by binary distinctions. Nevertheless, it seems still hard to establish an exhaustive list of what we call proper nouns. Actually, proper nouns, important in number and in frequency, are one of the most problematic units for all automatic analyzers of natural language texts.

In this study, we will focus on the problems of recognition of proper **names**. We do not try to characterize them as an axiomatic class, but attach to them a formal definition to determine explicitly the object of our study. Here is our formal criterion:

{χ ∈ (Interrogative Pron 누구? [Who?]) | χ ∉ (DECOS-NS) }



That is, proper names are determined by the fact that they do not exist in our lexicon of Korean common nouns (DECOS-NS/V01) [Nam94], and by their correspondence with the interrogative pronoun '누구 *nugu?* [who?]'. The nouns considered as proper names according to these conditions do not always correspond to our semantic intuition. Nevertheless, they usually do not have intrinsic meanings; and they do not have explicitly distinct referents. Given that a lexicon of Korean common nouns (DECOS-NS) has already been built [Nam94], the ambiguity across the category of common nouns and that of proper ones will be settled only in one of these two lexicons by looking up DECOS-NS: if they already are included in this lexicon, we do not consider them in the lexicon of proper nouns, without questioning their linguistic status. Remember that our goal is not to discuss the nature of this noun class, but to complete the lexicon of Korean Nouns in NLP systems. In order to handle them in an NLP system, given that we do not have yet a dictionary which provides all proper nouns, auxiliary methods are required, such as syntactic information or local grammars that allow to analyze them.

In the following sections, we will classify in five types the contexts where Proper Names can appear, and describe their characteristics in detail.

2. Typology of PN Contexts

2.1. Type I < *PN-(Postposition + E)* >

This type of noun phrases is without any particular characteristics inherent to *Proper Names* (*PN*). They actually occur in the positions of common nouns, as shown in the following graph (Figure 2):

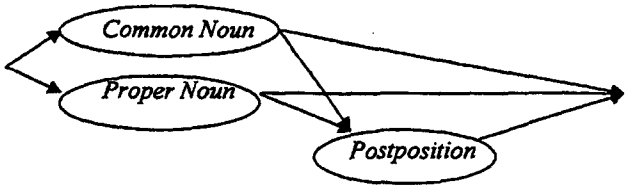


Figure 2. Type I of Nominal Phrases containing *PNs*

Postpositions observed on the right side of nouns (proper or common ones) indicate grammatical functions of the attached noun. When they appear in this context, there are no ways to sort out proper names, only by analyzing their syntactic structures. Let us consider:

김정일이 북한의 통치자이다
Kim Jung Il - i bughan-eui tongchija-ida
PN<Kim Jung Il>-Postp North Korea-of president-be
(Kim Jung Il is the President of North Korea)

We cannot distinguish this *PN* <Kim Jung Il> from other nouns that can be found in this position, such as in the following:

(그 남자 + 조선인)-이 북한의 통치자이다
(geu namja + josenin)-i bughan-eui tongchija-ida
((This man +A Korean) is the President of North Korea)

As mentioned above, in English or in French, proper names could be distinguished from common nouns, at least by means of the use of the upper case for the former, even though it is not an absolute criterion. Consider:

Jacques Chirac est le Président de la France
Bill Clinton is the President of USA

Nevertheless, the upper case does not totally satisfy our semantic intuition, since we also observe nouns with the upper case, such as *Président* or *President*, which certainly do not designate one particular person (here, we encounter the fundamental problem of the definition of the term 'proper'). Likewise, in the following sentence, the noun *Français* and *American* started with the upper case cannot be considered as proper names, whatever the definition of proper name is:

(Cet homme + Un Français) est le Président de la France
(This man + An American) is the President of USA

2.2. Type II < *PN (Spec+E) Professional Title-(Postposition+E)* >

This type of sequence is characterized by the presence of nouns of *professional title* (*PT*), such as:

박사	*bagsa*	[Doctor]
교수	*gyosu*	[Professor]
원장	*wenjang*	[Director]
사장	*sajang*	[President]
장관	*janggwan*	[Minister]

For example:

(1)김 민우 박사는 미국에서 5년간 공부했다
Kim MinU bagsa-neun migug-eise 5nyengan gongbuha-essda
PN<Kim MinU> Doctor-Postp U.S.A.-Postp 5 year-during study-Past
(Dr. Kim MinU has studied in U.S.A. during 5 years)

The noun phrase in subject position: *'Kim_ MinU_ bagsaneun'* is composed of three strings. However, in Korean, typographical constraint is not a reliable criterion, since we cannot prohibit writing this phrase in other ways like:

(2a)	김민우 박사는	*KimMinU_ bagsaneun*
(2b)	김민우박사는	*KimMinUbagsaneun*

When proper names occur as attached to other elements of noun phrases, their analysis becomes more complicated. Therefore, a local grammar recognizing *PTs* such as (Figure 3):

Figure 3. Local grammar of Type II

will reduce numerous mismatchings between the strings like (2b) and the combination of the items found in a dictionary.

Since a family name alone can precede *PTs*, the grammar above should be refined (Figure 4):

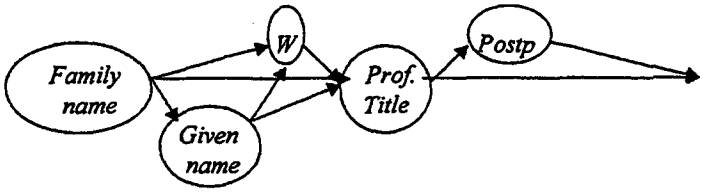


Figure 4. A more detailed Local grammar of Type II

Thus we observe (3) instead of (1):

(3) 김 박사는 미국에서 5년간 공부했다
Kim bagsa-neun migug-eise 5 nyengan gongbuha-essda
(Dr. Kim has studied in the U.S.A. during 5 years)

while a given name alone hardly appears with *PTs*:

??민우 박사는 미국에서 5년간 공부했다
??MinU bagsa-neun migug-eise 5 nyengan gongbuha-essda
(Dr. Min U has studied in U.S.A. during 5 years)

When we list the nouns of professional title, the number of *PNs* recognized by the local grammar presented in Figure 4 will be increased. Nevertheless, listing these nouns up does not guarantee automatically to recognize *PNs*, since we can come across specific nouns (*Spec*) inside of these sequences:

(4) 김 민우 법학 박사는 미국에서 5년간 공부했다
Kim MinU bebhag bagsa-neun migug-eise 5nyengan gongbuha-essda
(Dr. of Laws Kim MinU has studied in U.S.A. during 5 years)

The *Specs* are appropriate to *PTs*: we observe nouns designating scientific domains such as 'physics', 'biology', 'mathematics', or 'literature' for the *PTs* like 'doctor', whereas we find another set of Specs for the *PT* 'minister': 'education', 'culture', or 'transfortation' for example.

Notice that *PTs* can also appear without *PNs*:

(법학 + E) 박사가 제일 선망받는 사회적 지위중의 하나이다
(bebhag + E) bagsa-ga jeil senmangbad-neun sahoijeg jiwijung-eui hana-i-da
(A doctor (of Law + E) is one of the most envied social titles)

그 (법학 + E) 박사는 일찍 한국을 떠났다
geu (bebhag + E) bagsa-neun iljjig hangug-eul ddena-ssda
(This doctor (of Law + E) left Korea early)

Thus, in order to analyze the strings followed by a *PT* in contexts such as (5), the system should first look up a lexicon of Common Nouns (and eventually a lexicon of Determiners), and if the search fails, one could suppose that we found a proper name:

(5a)이공계 박사가 인기가 높다
igonggyei bagsa-ga ingi-ga nop-da (Doctors of Natural Science are highly requested)

(5b)이 공학 박사가 인기가 높다
i gonghag bagsa-ga ingi-ga nop-da (This doctor of Science is highly requested)

(5c)이민우 박사가 인기가 높다
iminu bagsa-ga ingi-ga nop-da (Doctor Lee MinU is highly requested)

In (5a), the string found with '*bagsa* [doctor]' is a simple noun '*igonggyei* [natural science]'; the sequence that precedes '*bagsa* [doctor]' in (5b) is a phrase composed of a determiner '*i* [this]' and a common noun '*gonghag* [science]'; the element followed by '*bagsa* [doctor]' in (5c) will not be matched with any entries of the lexicon of common nouns: only this string will then be recognized as a proper name.

The local grammar proposed so far should be completed by the description of the following transformation. Let us compare (4) with (6):

(6)법학 박사 김 민우는 미국에서 5년간 공부했다
bebhag bagsa Kim MinU-neun migug-eise 5 nyengan gongbuha-essda
(Kim MinU, Dr. of Laws, has studied in U.S.A during 5 years)

The sentence (6) can still be transformed into:

김 민우는 미국에서 5년간 공부한 법학 박사이다
Kim MinU-neun migug-eise 5 nyengan gongbuha-n bebhag bagsa-ida
(Kim MinU is a doctor of Laws who has studied in U.S.A. during 5 years)

In fact, the sequence containing *PTs* corresponds to a simple sentence:

PN W-Professional Title
= *W-Professional Title PN*
= *S: PN be a W-Professional Title*

(7a)	김 민우 법학 박사	*Kim MinU bebhag bagsa*	[Dr. of Law Kim MinU]
(7b)=	법학 박사 김 민우	*bebhag bagsa Kim MinU*	[Kim MinU, Dr. of Law,]
(7c)=	김 민우는 법학 박사이다	*Kim MinU-neun bebhag bagsa-ida*	[KimMinU is a Dr. of Law]

2.3. Type III <*PN-(Gen+E) Family Relation-(Postposition+E)* >

This type of phrases contains nouns designating a *family relation* (*FR*) such as:

아들	*adeul*	[son]
아버지	*abeji*	[father]
처	*che*	[wife]
손자	*sonja*	[grandchild]
며느리	*myeneuli*	[daughter-in-law]

These nouns have a strong possibility to occur with a proper name, as shown in the following sentence:

김민우의 아들은 올해 20 살이다
Kim Min U-eui adeul-eun olhai 20 sal-ida
PN<Kim Min U>-Gen son-Postp this year 20 years old-be
(Kim Min U's son is 20 years-old this year)

The Genitive Postposition '의 *eui* ['s/of]' can be omitted:

김 민우 아들은 올해 20 살이다
Kim MinU adeul-eun olhai 20 sal-ida
(Kim Min U son [='s son] is 20 years-old this year)

The structure can be formalized in the following graph (Figure 5):

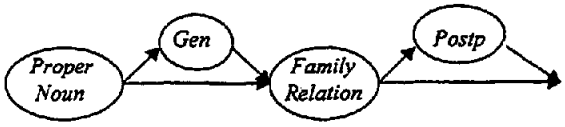


Figure 5. Type III of noun phrases containing *PNs*

The strings *'N-(Gen+E) FR'* do not automatically guarantee existence of proper names, since common nouns that have a human feather can also appear with a *FR* like:

(빵집 주인 + 옆집 남자) - (의 + E) 아들은 올해 20 살이다
(bbangjib juin+yepjib namja)-(eui+E) adeul-eun olhai 20 sal-ida
((The baker + The neighbor)'s son is 20 years-old this year)

In fact, strings containing *FRs* are necessarily based upon human nouns, proper names being only one class of human nouns. This context helps to find proper names, but is not a sufficient condition to recognize them automatically.

2.4. Type IV *<PN Vocative Term-(Postposition+E) >*

We call *Vocative Terms* (*VT*) the following utterances:

영감 !	*yenggam !*	[Sir !]
선배님!	*senbainim !*	[Senior !]
누나 !	*nuna !*	[Elder sister ! (for a boy)]
언니 !	*enni !*	[Elder sister ! (for a girl)]
형 !	*hyeng !*	[Elder brother ! (for a boy)]
오빠 !	*obba !*	[Elder brother ! (for a girl)]

The nouns above can all be used as *VTs*, that is, a term one can use to indicate some social or familial relations between himself (i.e. the speaker) and his interlocutor(s), or to call on somebody paying due respect to his social status (honorific terms). In addition, with proper names, they can also occur in assertive sentences, like:

김 영감이 왔다
Kim yenggam-i wa-ssda PN<Kim> sir-Postp come-Past (Sir. Kim came)

인아 누나가 떠났다
Ina nuna-ga ddena-ssda PN<In A> elder sister-Postp leave-Past (Elder Sister InA left)

These *VTs* should be compared with the nouns of professional title (*PT*) that we examined in section 2.2. and those of family relation (*FR*) mentioned in 2.3., since some of them (*PTs* and *FRs*) can also be used in calling someone, like in:

김교수 !	*Kim Gyosu!* (Come here !)	[PT: Professor Kim !]
김민우 원장님 !	*Kim MinU Wenjangnim !*	[PT: Director Kim MinU !]

아버지 !	*abeji !*	[FR: Father !]
어머니 !	*emeni !*	[FR: Mother !]

Let us examine differences among them:

♣ Difference between *VT* and *PT*

Nouns of Professional Title (*PT*) are different from Vocative terms (*VT*), not only in syntactic, but also in semantic ways. As *PTs* do not have inherently vocative functions, they can hardly be used alone in the vocative case:

?*교수 ! *?*gyosu !* [Professor !]
?*장관 ! *?*janggwan !* [Minister !]

Then, one should either attach to them a vocative suffix such as '님 *nim*', or adjoin them to proper names:

교수님 ! / 김 민우 교수 ! *gyosu-nim!/Kim MinU gyosu!*
장관님 ! / 김장관 ! *janggwan-nim!/Kim janggwan!*

Semantically, *PTs* designate professions, the list of which we can determine a priori, while *VTs* are more vague and non-predictable without examining pragmatic situations: the latter are closer to the nouns of Family Relation (*FR*), since, as mentioned above, they imply familial or social relations between a speaker and his interlocutor(s).

♣ Difference between *VT* and *FR*

What we call nouns of Family Relation (*FR*) cannot appear with a proper name when they are used in the vocative case. Thus, is not allowed the internal structure:

**ProperName FamilyName !*

such as:

(1) *김민우 아버지 ! **KimMinU abeji!* [FR:Father Kim MinU!]
*박어머니 ! **Park emeni !* [FR: Mother Park !]

Remember that *FRs* are formally defined as occurring in the structure *'PN-Gen FR* [Proper Name's FR]', thus, when we encounter them in the contexts *'PN FR'* (e.g. 김민우 아버지 *Kim MinU abeji* [Kim MinU]), such strings have a meaning corresponding to *'PN-Gen FR'* (e.g. 김민우의 아버지 *Kim MinU-eui abeji* [Kim MinU's father]): *PNs* are not appositions to *FR*, like in sequences composed of *'PN VT'* such as (2). *VTs*, by definition, should be able to appear directly associated with proper names. Compare (1) with:

(2) 김민우 형 ! *Kim MinU hyeng !* [VT: Brother Kim MinU !]
박형 ! *Park hyeng !* [VT: Colleague Park !]
박영감 ! *Park yenggam* [VT: Sir Park !]
민우 오빠 ! *Minu obba !* [VT: Brother MinU !]

Let us underline that some *VTs* do not accept family names alone, whereas some others allow them, as well as given names alone or full names. Here are some cases (Figure 6):

	F.name alone	G.name alone	Full name
형1 *hyeng1*	-	+	+
형 2 *hyeng2*	+	-	-
오빠 *obba*	-	+	+
영감 *yenggam*	+	-	+
선배 *senbai*	+	+	+

Figure 6. Some *VTs* with their associated *PN* types

2.5. Type V *<PN Incomplete Noun-(Postposition+E)>*

This type of Noun Phrase is similar to the preceding one: what we call *Incomplete Nouns* (*IN*) is also used for social appellation. However, they are different from the preceding ones by the fact that they do not have syntactic autonomy, and therefore they never can appear alone in any positions of a sentence. Here is their list:

씨 *ssi* [Mr. / Miss. / Mrs.]
양 *yang* [Miss.]
가 *ga* [Mr. <pejorative>]
님 *nim* [Mr. / Miss. / Mrs. <respectful>]
군 *gun* [Mr.<young boy>]
옹 *ong* [Mr. <old man>]

Let us consider:

김 민우씨가 왔다 *Kim MinU - ssi - ga wa-ssda*
PN<Kim MinU> - IN[Mr.] - Postp come-Past (Mr. Kim MinU came)

김양이 떠났다 *Kim - yang - i ddena-ssda*
PN<Kim> - IN[Miss.] - Postp leave-Past (Miss. Kim left)

Notice that *PNs* vary according to *INs*. The following table represents different types (Figure 7):

	F. Name	G. Name	Full Name
씨 *ssi*	+	+	+
양 *yang*	+	+	+
가 *ga*	+	-	-
님 *nim*	-	-	+
군 *gun*	+	+	+
옹 *ong*	+	-	+

Figure 7. Types of *PNs* according to *INs*

The table above can be represented by a Finite State Automaton (FSA) [Lap96], [Gro87] as shown in the following graph (Figure 8):

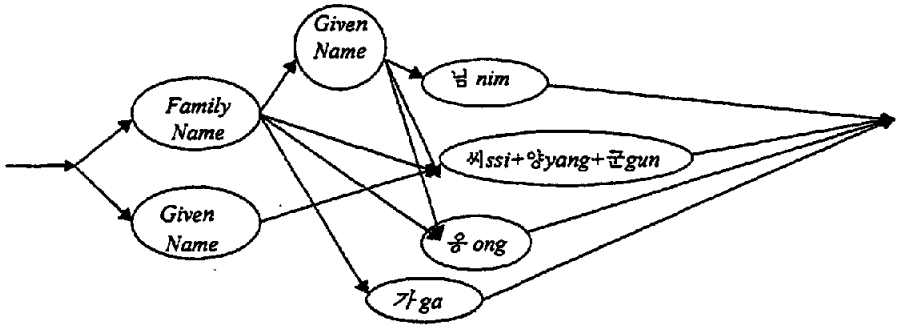


Figure 8. FSA of *PN-IN*

These nouns (*INs*), syntactically and semantically incomplete, always require proper names to their left side. In this sense, this type of contexts is appropriate to *PNs*: if an *IN* is recognized, we can be assured to find a *PN* near to it. In spite of this strong constraint, since all *INs* are mono-syllabic, ambiguity problems are often hard to handle. For example, the *IN* '가 *ga* [Mr.]' is an homograph of several items. Let us consider some of them (Figure 9):

Type	Part of Speech	*Meaning*	Example
가*ga*	Incomplete N.	*Mr.*	박가
	Simple Noun	*grade*	수우미양가
	Prefix1	*dance*	가곡
	Prefix2	*provisory*	가석방
	Suffix1	*letter*	병가

	Suffix2	*value*	영양가
	Suffix3	*family*	처가
	Suffix4	*music*	유행가
	Suffix5	*person*	무용가
	Suffix6	*boundary*	화롯가
	Suffix7	*area*	대학가
	Verb	*go*	떠나가 버렸다
	Postposition	*Nominative*	우리가
	Terminal Sfx	*Interrogation*	떠났는가?

Figure 9. Homograph types of '*ga*'

The following sentence illustrates this ambiguity problem:

(1) 박가 친구들과 우리가 무허가 주택가 근처 한 우물가에서 유행가를 부르고 있을 때, 영양가 없는 빵부스러기 주위에 몇 마리 새가 앉아 있었던가?

bag-ga chingu-deul-gwa uli-ga muhega jutaig-ga geunche han umul-ga-eise yuhaing-ga-leul buleu-go isseul ddai, yengyang-ga ebs-neun bbangbuseulegi juwi-ei myech mali sai-ga anja iss-ess-den-ga ?

(When we were singing popular songs with Mr. Park's friends at the edge of a well near the area of unlicensed buildings, how many birds were there sitting around bread crumbs without any taste ?)

We observe the morpheme *ga* 9 times. But only the first occurrence of *ga* is an Incomplete Noun which accompanies a *PN*. In the 8 other strings, we should not expect occurrences of *PNs*: in order to recognize an *IN ga*, first, dictionaries of all common nouns (i.e. simple nouns, derived nouns, and compound nouns) must be available. If the string containing *ga* is not found in these dictionaries, then the final syllable *ga* might be a verb, a nominative postposition attached to a noun, or an inflectional suffix attached to a verb; or else, it is an IN *ga*.

In the case of (1), strings containing *ga*, such as the following ones, are detected as common nouns (simple or derived ones):

무허가	*muhega*	unlicensed
주택가	*jutaig-ga*	area of buildings
우물가	*umul-ga*	edge of a well
유행가	*yuhaing-ga*	popular songs
영양가	*yengyang-ga*	any taste

and the following ones are either nouns followed by a postposition *ga* or a verb including the inflectional suffix (*IS*) *ga*:

우리가	*uli-ga*	we-Postp
새가	*sai-ga*	bird-Postp
있었던가	*iss-ess-den-ga*	be-IS [Past-Past-Interrogation]

The string '박가 *bag-ga*' will not be recognized as one of these cases, even though there exists a simple noun '*bag* [pumpkin]' in the dictionary of common nouns, since the postposition required by this noun is not '가 *ga*', but '이 *i*'. Therefore, *bag-ga* will be analyzed as a proper name *bag* (family name alone) followed by an *IN ga*.

3. Building Local Grammars of *PNs*

Let us summarize the formal definition of the five contexts where a Proper Name (*PN*) can occur:

♣ Type I. Noun Position : *<PN-Postp>*
♣ Type II. With Professional Title(PT) : *<PN-(Spec)-PT-Postp>*
♣ Type III. With Family Relation (FR) : *<PN-Gen-FR-Postp>*
♣ Type IV. With Vocative Term (VT) : *<PN-VT-Postp>*
♣ Type V. With Incomplete Noun (IN) : *<PN-IN-Postp>*

These five contexts are represented in Figure 10:

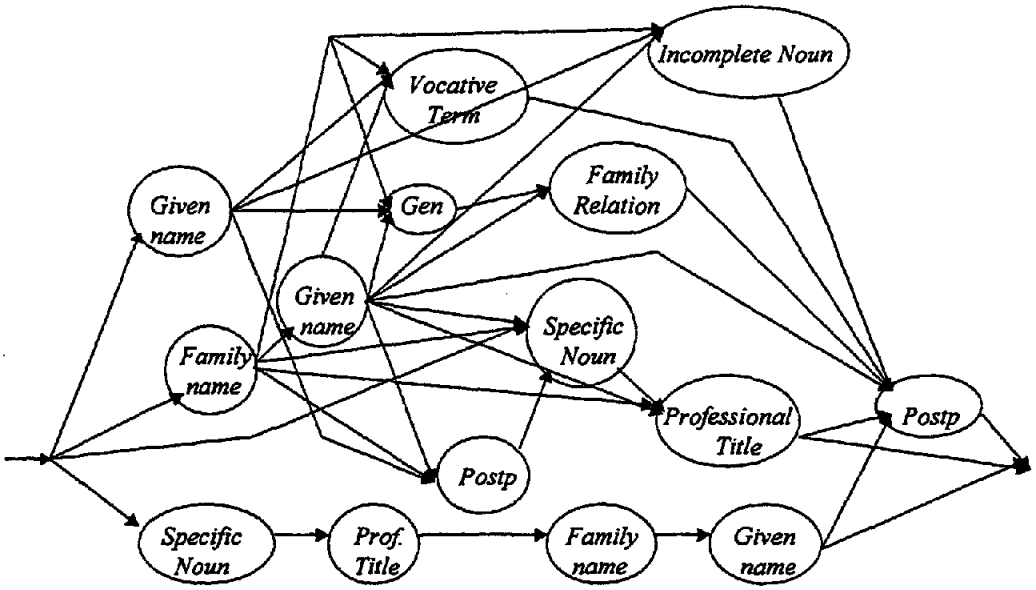


Figure 10. Local grammar of *PNs*

Notice that when we recognize Incomplete Nouns (i.e. 씨 *ssi*, 양 *yang*, 가 *ga*, 님 *nim*, 군 *gun*, 옹 *ong*), the occurrence of proper names is guaranteed, since *INs* cannot occur without *PNs*. Nevertheless, as mentioned above, serious ambiguity problems appear in the distinction of *INs* from their homographs, we here propose two complex local grammars in order to increase the ratio of identification of *INs*.

3.1. Use of *PostHN* appropriate to Human Nouns

There are specific items appropriate to human nouns: we name them *PostHN*. They do not constitute autonomous units, but are attached to human nouns at the syntactic level. Thus, they appear even after the plural marker 들 *deul* [/s/]. For example, in the following sentences, a *PostHN* '네 *nei* ['s family/house]' appears with a *PN* alone, or with a *PN* followed by an *IN* (here, 씨 *ssi* [Mr.]):

민우네가 이 마을에서 제일 부지런하다 *MinU-nei-ga i maeul-eise jeiil bujilenha-da*
PN<MinU>-PostHN[family]-Postp this village-Postp most diligent-St
(MinU's family is most diligent in this village)

강진오씨네에서 불이 났다 *Gang GinO-ssi-nei-eise bul-i na-ssda*
PN<Kang GinO>-IS[Mr.]-PostHN[house]-Postp fire-Postp occur-Past
(There was a fire in Mr. Kang GinO's house)

In French, we observe a preposition similar to this *PostHN*: *chez* ('s family/house), a locative preposition, as *at one's* in English, which selects only human nouns:

Il y a eu un feu chez M. Pierre Picon
There was a fire at M. Pierre Picon's

Therefore, when we encounter a sequence that ends with an *IN-PostHN-Postp,* the possibility to find a *PN* is increased. For example, the following string:

장가네는 *jang-ga-nei-neun*

can be analyzed in 510 ways (i.e. (7 x 7 x 5 x 2) + (2 x 5 x 2) = 510) after a simple matching of the words of this string with their lexicon entries (Figure 11):

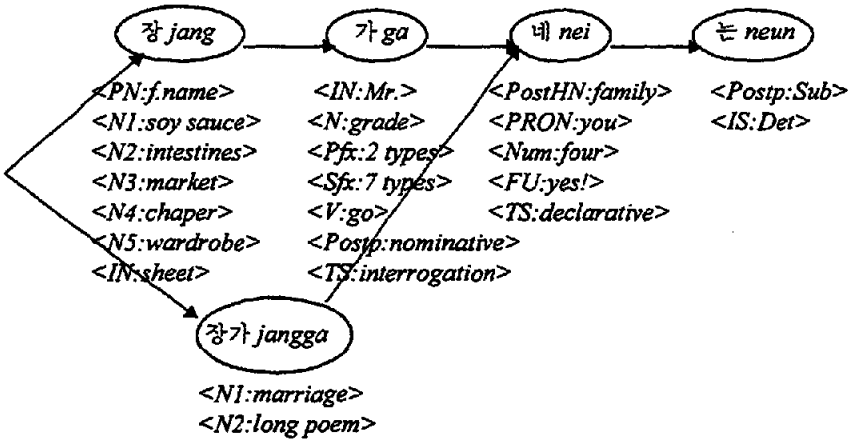


Figure 11. 510 analyses of '장가네는 *jang-ga-nei-neun*'

According to the local grammars we have constructed, we get the following result for this string (Figure 12):

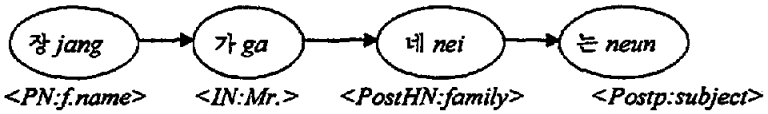


Figure 12. Accurate analysis of the string in Figure 11

3.2. Superposition of contexts for *PNs*

Let us examine the following sentences:

(1a)김씨동생네집은 아주 크다
Kim-ssi-dongsaing-nei-jib-eun aju keu-da
PN<Kim>-IN-brother-PostHN-appartment-Postp very large-St
(The appartment of Mr.Kim's brother's family is very large)

(1b)김민우박사님 아들네에서 잔치가 열렸다
Kim MinU bagsa-nim-adeul-nei-eise janchi-ga yelli-essda
PN<Kim MinU>-doctor-Sfx-son-PostHN-Postp party-Postp occur-Past
(There was a party in Dr. Kim MinU's son's family's house)

(1c)진오형님네로 모두 가자!
GinO hyeng-nim-nei-lo modu ga-ja!
PN<GinO>-brother-Sfx-PostHN-Postp together go-St
(Let's go together to Brother GinO's house !)

Here, several of the noun phrases we have examined so far occur piled together. The internal structures of the examples above are respectively:

(2a) *PN - <Type V> - <Type III> - PostHN - Noun - Postp*
(2b) *PN - <Type II> - <Type III> - PostHN - Postp*
(2c) *PN - <Type IV> - PostHN - Postp*

Hence, by providing information about the combinations of these strings, we could rise the accuracy in recognizing *PNs*. For example, the string that includes the sequence 씨동생네집 *ssi-dongsaing-nei-jib* in (1a) can hardly be anything else than a noun phrase containing a *PN*. Thus, even though we find several entries 김 *kim* in the lexicon of nouns, such as:

김 *kim*	1. Noun = steam	[e.g. 김이 난다]
	2. Noun = dried laver	[e.g. 김밥]
	3. Noun = hope	[e.g. 김이 샜다]
	4. Completive Noun = chance	[e.g. 온 김에...]

we can eliminate these interpretations, since these forms precede the complex sequence that requires necessarily a *PN*.

4. Experimental results

So far, we have examined contexts where we expect to encounter Proper Names (*PN*). In order to recognize automatically *PNs* on a large scale in texts in the absence of a complete lexicon of *PNs*, the description of noun phrases containing *PNs* should be necessary. We constructed local grammars based upon our description of the types of nominal phrases containing proper names.

Notice that implementing such a system requires the use of the relation between **Recall** and **Precision**. In general, it is understood that Recall is the ratio of relevant documents retrieved for a given query over the number of relevant documents for that query in a database, and Precision is the ratio of the number of relevant documents retrieved over the total number of documents retrieved [Fra92].

However, Recall-Precision plots show that Recall and Precision are inversely related. That is, when Precision goes up, Recall typically goes down and vice-versa. If we want to recognize automatically *PNs* in a given text in order to construct an electronic lexicon of *PNs*, Recall, that is the ratio of *PNs* retrieved for a given grammar over the number of *PNs* in the text, should certainly be higher than Precision.

Let us consider some experimental results of our study. In the contexts of Type V, i.e. *<PN Incomplete Noun-(Postposition + E)>*, the Incomplete Noun (*IN*) '씨 *ssi* [Mr./ Miss./ Mrs.]' can appear with a family name alone, a given name alone, or a full name (*cf.* 2.5. Figure 7). Remember that, in Korean, a typographical unit delimited by blanks cannot directly be taken as a basic element for morphological analysis [Nam97]: we should then analyze the strings occurring with a blank on the left side of *INs* as well as the strings stuck to *INs* in order to examine the context Type V. Thus, the local grammar of Type V for '씨 *ssi*' is the following graph (Figure 13):

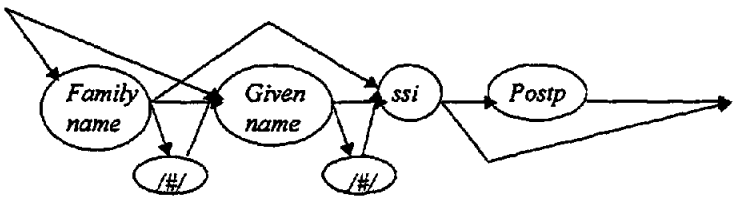


Figure 13

Our first text was composed of 29373 characters [Cho96], we located 22 sequences containing *ssi*, 20 of which are *PNs* (Figure 14):

F.Name *-ssi*	G.Name *-ssi*	Full Name *-ssi*	non*PN*	Total
12	0	8	2	22

Figure 14

In the second text, composed of 30869 characters, 69 occurrences of *'X-ssi'* are observed. All *'Full name-ssi'* sequences here appear attached, whereas, in the preceding text, they all appear with a blank (i.e. *'X-#-ssi'*). Here is the result (Figure 15):

F.Name *-ssi*	G.Name *-ssi*	Full Name *- ssi*	non*PN*	Total
35	1	24	9	69

Figure 15

The 9 sequences *'non-PNs'* are as followings:

[1] 억씨기	*eg-ssi-gi*	[6] 아가씨,	*aga-ssi,*
[2] 입씨름을	*ib-ssi-leumeul*	[7] 제씨되는	*jei-ssi-doineun*
[3] 아가씨란	*aga-ssi-lan*	[8] 날씨였다	*nal-ssi-yessda*
[4] 아가씨두	*aga-ssi-du*	[9] 날씨가	*nal-ssi-ga*
[5] 아가씨라	*aga-ssi-la*		

Looking up our dictionaries of Korean Simple Nouns (DECOS-NS/V01) [Nam94], and of Korean Postpositions (DECOS-POST/V01) [Nam96b] eliminate [2], [3], [4], [5], [6], [8], [9], which are the sequences composed of a common noun and a postposition (or a typographical separator such as a comma). Because [1] is a dialectal adverb, and [7] a 'Noun-Verb' string, they were not detected in our system.

The third text composed of 33982 characters contains 10 occurrences of *'X-ssi'* one of which is a *nonPN* ('프로씨름이 *peulo ssi-leumi* [N/N/Postp]') (Figure 16):

F.Name - *ssi*	G.Name - *ssi*	Full Name *-ssi*	non*PN*	Total
4	0	5	1	10

Figure 16

This *nonPN* was eliminated after looking up the dictionary of Postpositions: there is no postposition '름이 *leumi*'. The analysis of the text above on the basis of the local grammar presented in Figure 3 (Type II *<PN Spec-PT-(Postp + E)>*) in 2.2. allows to recognize *PNs* in a more satisfactory way. Besides *'X-ssi'* strings, with two *PTs*: '대통령 *daitonglyeng* [the President]' and '수상 *susang* [the prime minister]', we could recognize 73 % of *PNs*, that is, 49 occurrences of 67 (i.e. Recall is 0.73). However, use of the local grammars of Figure 13 and Figure 3 (only with these two *PTs* above) leaves some *nonPNs*: Precision is 0.7 (49 strings of 70 which occurred with these *IN* and *PTs* are *PNs*). Since our goal is to recognize most contexts where *PNs* can occur, in order to construct a lexicon of *PNs* as complete as possible, Recall should be more important than Precision in our system. By adding a few more *PTs* (cf. Type II) such as '장군 *janggun* [general]', '선수 *sensu* [player]', *FRs* (cf. Type III) such as '부녀 *bunye* [father-daughter]', or *INs* (cf. Type V) such as '양 *yang* [Miss.]' in the lexicon on the basis of which our local grammars are constructed, we could obtain a more reliable result as shown in the following table (Figure 17):

Retrieved strings	*PNs* retrieved	Existing *PNs*
89	59	67

Figure 17

Thus, Recall increases: 0.88, whereas Precision goes down: 0.66. The 8 *PNs* that are not retrieved by our local grammars are given below. Actually, their contexts are hard to determine, since they are syntactically identical with contexts where common nouns can appear:

[1] '사또오',	*saddoo*	[5] '김대중 사건',	*kimdaijung sagen*
[2] '맥아더'의	*maigadeeui*	[6] 문세광 사건	*munseigwang sagen*
[3] '도요또미'의	*doyoddomieui*	[7] 김지하나	*kimjihana*
[4] 김대중 문제라든가	*kimdaijung munjeiladeunga*	[8] 김대중 사건,	*kimdaijung sagen*

To guarantee that all occurrences of *PNs* are covered by local grammars, it would be necessary to consider a great part of the contexts where common nouns appear.

In this paper, we have described the contexts where proper names can occur, but the complete lists of the nouns requiring *PNs* have not been done. We are sure that these lists are not illimited ones, they will be presented in further studies. Notice that these studies are deeply related to the syntax of nouns, especially that of human nouns. In this sense, human noun, a semantic concept, can nonetheless become an operational term in the formal description of natural languages, indispensable many procedures of Natural Language Processing (NLP) systems.

References

[Cha93]Chang, Chao-Huang, 1993, Corpus-based Adaptation Mechanisms for Chinese Homophone Disambiguation, Proceedings of the Workshop on Very Large Corpora, Ohio State University, USA.

[Cho94]Choi, Key-Sun et al, 1994, A Two-Level Morphological Analysis of Korean, Proceedings of the 15th International Conf. on Computational Linguistics (COLING '94), Kyoto, Japan.

[Cho96]Choi, Key-Sun et al, 1996, Korean Information Base Corpus, KAIST.

[Cou87]Courtois, Blandine, 1987, Dictionnaire électronique du LADL pour les mots simples du français (DELAS), Rapport Technique du LADL, N°17, University Paris 7.

Dictionnaire universel des noms propres (Petit Robert 2), 1974, ed. Le Robert, Paris, 1st edition.

[Fra92]Frakes, William B.; Ricardo Baeza-Yates, 1992, Information Retrieval: Data Structures and Algorithms, Prentice Hall, Englewood Cliffs, New Jersey 07632.

[Gar91]Gary-Prieur, Marie-Noëlle, 1991, Le nom propre constitue-t-il une catégorie linguistique?, Langue française N-92, Paris: Larousse.

[Gro87]Gross, Maurice, 1987, The use of finite automata in the lexical representation of natural language, Lecture Notes in Computer Science 377, Springer-Verlag.

[Gro89]Gross, Maurice, 1989, La construction de dictionnaires électroniques, Annales des Télécommunications, tome 44 N°1:2, Issy-les-Moulineaux-Lannion: CNET.

[Gro93]Gross, Maurice, 1993, Lexicon Based Algorithms for the Automatic Analysis of Natural Language, in Theorie und Praxis des Lexikons, Walter de Gruyter: Berlin.

[Hee93]Heemskerk, Josee S., 1993, A probabilistic Context-free Grammar for Disambiguation in Morphological Parsing, Proceedings of the 6th Conference of the European Chapter of the Association for Computational Linguistics, Utrecht, The Netherlands.

[Lap96a]Laporte, Eric, 1996, Context-free parsing with finite-state transducers, RT-IGM 96-13, University of Marne-la-Vallée.

[Moh94]Mohri, Mehryar, 1994, Application of Local Grammars Automata: an Efficient Algorithm, RT-IGM 94-16, University of Marne-la-Vallée.

[Nam94]Nam, Jee-Sun, 1994, Dictionnaire des noms simples du coréen, RT N° 46, Laboratoire d'Automatique Documentaire et Linguistique, University Paris 7.

[Nam95]Nam, Jee-Sun, 1995, Constitution d'un lexique électronique des noms simples en coréen, Actes du LGC-1995 : Lexique-grammaires comparés et traitements automatiques, University of Québec à Montréal, Canada.

[Nam96a]Nam, Jee-Sun, 1996a, Dictionary of Korean simple verbs: DECOS-VS/01, RT N-49, LADL, University Paris 7.

[Nam96b]Nam, Jee-Sun, 1996b, Dictionary of Noun-Postpositions and Predicate-Postpositions in Korean: DECOS-PostN / DECOS-PostA / DECOS-PostV, RT N- 51, LADL, University Paris 7.

[Nam96c]Nam, Jee-Sun, 1996c, Construction of Korean electronic lexical system DECO, Papers in Computational Lexicography Complex '96, ed. by Ferenc Kiefer, Gabor Kiss et Julia Pajzs, Budapest : Linguistics Institute, Hungarian Academy of Sciences.

[Nam96d]Nam, Jee-Sun, 1996d, Classification syntaxique des constructions adjectivales en coréen, Amsterdam-Philadelphia: John Benjamins Publishing Company.

[Nam97]Nam, Jee-Sun, 1997, Lexique-Grammaire des adjectifs coréens et analyses syntaxiques automatiques, Langages N°124, Paris : Larousse.

[Nar93]Narayanan, Ajit; Lama Hashem, 1993, On Abstract Finite-State Morphology, Proceedings of the 6th Conference of the European Chapter of the Association for Computational Linguistics, Utrecht, The Netherlands.

[Oga93]Ogawa, Yasushi; A.Bessho; M.Hirose, 1993, Simple Word Strings as Compound Keywords: An Indexing and Ranking Method for Japanese Texts, Proceedings of the 16th Annual International ACM SIGIR, Pittsburgh, USA.

[Par94]Park, Se-Young et al., 1994, An Implementation of an Automatic Keyword Extraction System, Proceedings of Pacfic Rim International Conference on Artificial Intelligence '94, Beijing, Chine.

[Par96]Park, Se-Young et al, 1996, Korean Corpus- based on News papers, ETRI.

[Per95]Perrin, Dominique, 1989, Automates et algorithmes sur les mots, Annales des Télécommunications, tome 44 N 1:2, Issy-les-Moulineaux-Lannion: CNET.

[Rey74]Rey, Alain, 1974, Présentation du Petit Robert 2.

[Sil93]Silberztein, Max, 1993, Dictionnaires électro-niques et analyse automatique de textes, Le système INTEX, Paris: Masson.

[Ton93]Tong, Xiang; Chang-ning Huang; Cheng-ming Guo, 1993, Example-Based Sense Tagging of Running Chinese Text, Proceedings of the Workshop on Very Large Corpora, Ohio State University, USA.

[Wil91]Wilmet, Marc, 1991, Nom propre et ambiguïté, Langue française N° 92, Paris: Larousse.