Construction of the Korean electronic lexical system DECO

JEE-SUN NAM

ABSTRACT

We here discuss the method and the principles we have adopted in constructing the Korean electronic lexical system DECO. Given that existing editorial dictionaries are not reliable for this purpose, the use of a large corpus is required. However, even though the scale of the corpus is considerably extended, we can never ensure that all basic lexical items occur. Therefore, a combinatorial linguistic method based upon explicitly defined lexical categories is necessary to obtain all morphological sets related to a given basic form. The results of exhaustive description will be represented by finite state automata in our electronic lexicons.

I. EDITORIAL DICTIONARIES AND A LARGE CORPUS

Given that most printed texts are now available in electronic forms, accumulation of this type of information is considerable. Hence, the development and refinement of natural language processing (NLP) systems are incessantly required in order to archive these documents and offer requested information in a better way.

In the implementation of all kinds of NLP systems, the construction of electronic lexicons is elementary and indispensable: it is necessary to build up reliable electronic lexicons on the basis of coherent and explicit principles.

The methods that have been adopted so far for the construction of Korean on-line dictionaries can be summarized to the following two procedures:

1. Use of editorial dictionaries

One uses existing editorial dictionaries, that contain some morphological and grammatical information such as indication of parts of speech or derivational relations among lexical entries. However, existing dictionaries that, whether in printed forms or in electronic forms, are a priori

conceived for human users, are hardly disposable for this purpose. Thus, there are some problems, especially such as the followings:

1.1. Assignment of parts of speech

The assignment of parts of speech to lexical entries is not done in an explicit and coherent way. For example, as no formal criterion is given to distinguish verbs from adjectives, some items are considered as verbs in one dictionary, and as adjectives in another. Likewise, a great number of adjective roots are treated as nouns, whereas they do not have any lexical autonomy. Unless we do examine these problems, there would be no meaning in applying detailed syntactic rules based on high level grammars to the sets named *verbs*, *adjectives* or *nouns*.

1.2. Information about derivational relations

Information about derivational relations among lexical entries is not integrated in a systematic and exhaustive way. Thus, lists of verbs derived from adjectives by means of some suffixes (i.e. *Adj-Sfx = Verb*) are far from being complete. Affixed nouns and compound nouns are also selected without any coherency. This aspect is much less problematic for human users than for machines, because the former can *guess* the lack of information by reasoning by analogy. In fact, derivational and compositional information should be either all dropped out from a basic lexicon to be completed in a systematic way or all presented.

1.3. Encyclopedic entries

For practical purposes, the editorial dictionaries of Korean are not reliable, since they contain not only lexical entries (language dictionaries), but also encyclopedic entries such as proper nouns. For example, people's names, geographic places, historical events, artistic works, etc. are integrated in dictionaries as well as lexical items. Moreover, the number of proper nouns is incessantly increasing and it is difficult to establish their repertory. It is necessary to separate these two types of dictionaries from each other, so that we can complete them gradually.

2. Use of a large corpus

One uses a large corpus to establish on-line dictionaries. We can measure the frequency of lexical items, and then we can handle apart a great number of entries registered in editorial dictionaries that do not (or rarely) appear in texts. This point is not without importance in the case of Korean, since the number of archaic expressions is considerable in existing editorial dictionaries: this advantage is not mere nothing.

This procedure also allows one to process derived and compound forms more easily than using editorial dictionaries. Let us consider an example. The formation of some types of compound nouns is very productive and it would be very long to construct their complete list. The following compound nouns are of '*NounNoun*' type, one of the most productive types:

빨강색	*BbalgangSaig*	[red color]
영어책	*YengeChaig*	[English book]
칼자루	*KalJalu*	[knife handle]

We can obtain an interesting repertory of this type of nouns by using a large corpus. The fact that they usually appear without any typographical blank in them and that they are usually followed by a

grammatical **postposition** (i.e. a grammatical particle such as nominative, genitive, or accusative postposition, etc. The functions of postpositions correspond to those of **prepositions** in English such as *of, to, for, by,* etc.) makes their recognition easy: if we omit the right part of a string, which is identified as one of the postpositions, the left part might be a noun, whether a simple noun or a compound one.

Notice that the smallest units in automatic processing of texts can be words, morphemes or still something else. For example, in English or in French, this basic unit is usually a string cut up by two separators (e.g. blank, apostrophe or comma): it can be then named a **word**. Thus, when we observe sentences such as the following, we identify 5 units in each case:

John is in this cafe
Jean est dans ce café

In the case of Korean, the units delimited by separators are not on the same level as in the above cases: most grammatical markers are typographically attached to verbs (such as tense suffixes, modal suffixes, and so on..), and to nouns (such as nominative postpositions, genitive postpositions, and so on..). Thus, in the following sentence, we identify 4 units, but the nouns corresponding to *John* and *cafe* in English are suffixed with grammatical markers, **nominative** and **locative** respectively:

존은	그	카페에	있다	
Jon-eun	*geu*	*kapei-ei*	*iss-da*	
John-**nmtf**	this	cafe-**loc**	be-St	[= John is in this cafe]

Then, it does not make any sense to consider that strings cut up by two separators are the smallest units in Korean. If one imagines the number of *'Noun-Postp'* strings that can be obtained by the combination of more than ten thousand simple nouns and a thousand sets of postpositions (notice that several postpositions can be linked to a noun and that these combinations can be described in a local grammar), one will easily understand why these strings must not be taken as basic units. Therefore, the recognition of **nouns** in these strings is required priori to automatic analysis of texts.

Likewise, we can here use this procedure to list up "compound" nouns: recognition of postposition(s) will not be too complicated, since their list is much smaller than that of nouns; and then, elimination of these parts can provide a list of compound nouns.

However, the situation is not so simple. This method, i.e. constructing lexicons of nouns (not only simple nouns, but also compound ones) by means of recognizing postposition(s) and deleting them from strings requires considerable refinement, for the following reasons:

2.1. Absence of postposition

All nouns are not necessarily followed by (a) postposition(s). Here are two cases:

2.1.1. Dropping of postpositions

Postpositions can be dropped out in some contexts: if those in noun strings in the above example (i.e. '*John*-**nmtf**' et '*cafe*-**loc**') can hardly be omitted, they can easily disappear in the following sentence. Consider:

선생님	학교	가셨니 ?	
sensaingnim	*haggyo*	*gasyessni ?*	
teacher	school	went ?	[= Did the teacher go to school ?]

However, describing these conditions and predicting the dropping of postpositions are not easy. Moreover, in the above case, it should be difficult to distinguish **nouns** from **adverbs** without syntactic analyses, since adverbs are usually not suffixed with postpositions:

선생님	방금	학교	가셨니 ?	
sensaingnim	*banggeum*	*haggyo*	*gasyessni ?*	
teacher	**a while ago**	school	went?	[=Did teacher go to school a while ago?]

2.1.2.Compound nouns

Nouns that constitute compound forms can appear separate from each other (i.e. with blanks between them). Then, postpositions will only be found at the end of the last noun of the compound sequence. In the following sentence, the compound sequence is '*jayu seigei* [liberty world]':

그들은	자유	세계를	구현하였다	
geuteul-eun	*jayu*	*seigyei-leul*	*guhyenhayessda*	
they	**liberty**	**world-Acc**	achieved	[=They achieved liberty world]

Sometimes, spaces are obligatorily required inside compound nouns. The following example illustrates a compound sequence composed of 8 nouns:

자연	보호	운동	추진	위원회	결성	합의안	채택
jayen	*boho*	*undong*	*chujin*	*wiwenhoi*	*gyelseng*	*habeuian*	*chaitaig*

nature protection movement driving committee organization proposition acceptance
[**Acceptance** of the **proposition** of the **organization** of the **driving committee** of **Nature protection movement**]

Notice that we can link some of them as in *'NN N NN N N N'* or *'NNN NN NN N'*, but we can not write *'NNNNNNNN'* (symbol * indicates 'unacceptable sequence'):

*자연보호운동추진위원회결성합의안채택
**jayenbohoundongchujinwiwenhoigyelsenghabeuianchaitaig*
NatureProtectionMovementDrivingCommitteeOrganisationPropositionAcceptance

Therefore, recognizing nouns by eliminating (a) postposition(s) is no more a reliable method in this case, because we only observe (a) postposition(s) at the end of the eighth noun of this compound sequence:

국회는 자연 보호 운동 추진 위원회 결성 합의안 채택을 서둘렀다
gughoi-neun ***jayen boho undong chujin wiwenhoi gyelseng habeuian chaitaig-eul[Acc]*** *sedulessda*
[The National Assembly hastened [Acceptance of the proposition of the organization of the driving committee of Nature protection movement]]

2.2. Homograph

There are many cases where postpositions and the final morphemes of nouns are homographs. Let us consider an example:

우리가	무허가	주택가	근처를	배회할때,...
uliga	*muhega*	*jutaigga*	*geuncheu-leul*	*baihoihalddai,...*
We-**nmtf**	no-per**mit**	house-**area**	around-Acc	loiter-when

[When we are loitering around no-permit housing area, ...]

In this sentence, the first occurrence of '*ga*' is a nominative *postpostion* (i.e. **Noun-*nmtf***), whereas the second and the third ones are not: the second one is a part of the *noun* '*hega*' which a prefix '*mu*' is attached to (i.e. **Pfx-*Noun***); the third '*ga*' is a *suffix* which is attached to a noun '*jutaig*' (i.e. **Noun-*Sfx***). In other terms, only the first noun is linked to a postposition '*ga*'. Therefore, it would not be correct to automatically consider every final '*ga*' as a nominative postposition.

Here, we come across an important problem. Obviously, in order to make an appropriate analysis of the sentence above, we need a lexicon of nouns containing all these items, that is, not only simple nouns, but also affixed and compound nouns. However, let us recall that the lists of affixed and compound nouns we can obtain from editorial dictionaries are far from being complete and made up without any explicit principles. Nevertheless, the use of a large corpus to build these lists is not suitable either: we never can enumerate all sets of affixed and compound forms by using this procedure. Then, more refined linguistic studies about the mechanism of derivational relations among lexical items, based upon formal and coherent principles are required to build up a reliable on-line dictionary. Let us emphasize that linguistic descriptions can not easily be reduced to powerful general **syntactic rules**. We here have mentioned only some problems concerning *noun* sequences, but it is certain that one will come across such problems in other cases.

In the next paragraphs, we will present the method we have adopted and the principles of construction of the Korean electronic lexical system DECO.

II. THE KOREAN ELECTRONIC LEXICAL SYSTEM *DECO*

1. Lexicons of simple items *DECOS*, affixed items *DECOA* and compound items *DECOC*

The lexical system DECO is constructed not only by using existing editorial dictionaries and a large Korean corpus, but also a combinatorial method based upon explicitly defined lexical categories.

First of all, all simple items are separated from complex forms on the basis of syntactic criteria. Thus, '여기자 *yegija* [woman journalist]' is a complex form, i.e. '**pfx**(*ye*)-**noun**(*gija*)', whereas '여자 *yeja* [woman]' is a simple noun, because, even though it also contains the initial morpheme *ye*, the other part *ja* is not an autonomous unit. Diachronic and semantic analogies are not considered, but syntactic properties are taken as classifying criteria.

We have classified all simple items in 5 types of parts of speech: *Nouns*, *Adjectives*, *Verbs*, *Adverbs*, and *Functional Units*. They are encoded as NS, ADJS, VS, ADVS, and FUS where S stands for *simple*. Some syntactic and morphological information is integrated in the form of codes such as **PRED1** that indicates 'nouns that can be accompanied by 하다 *Hada* to form a sequence equal to a transitive verb' or **SM** that means 'adjectives the ending form of which is 스럽다 *Seulebda*. Each category itself is divided into sub-categories. These simple items constitute the lexicon DECOS (Korean electronic dictionary of simple items). The number of entries in the current version [DECOS-V01] is shown in the following table <figure 1>:

Nouns	*Adjectives*	*Verbs*	*Adverbs*	*Functional Units*	Total
15 000	5 300	7 500	7 000	200	35 000

< figure 1- number of the entries of DECOS-V01 >

Here are the first entries of the lexicon of simple nouns [DECOS-NS / V01] <figure 2> and those of the lexicon of simple adjectives [DECOS-ADJS / V01] <figure 3>:

가 NS.	가늠 NS. /PRED1	가르매 NS. /NVS/*PREDHA	가물치 NS. /ANM
가간 NS.	가다랭이 NS.	가리 NS.	가뭄 NS. /NVM
가감 NS. /PRED1/PRED3	가닥 NS.	가리개 NS. /NVS/*PREDHA	가미 NS. /PRED1/PRED3
가객 NS. /HUM	가대인 NS. /HUM	가리마 NS. /NVS/*PREDHA	가발 NS. /*PREDHA
가게 NS. /*PREDHA	가도 NS.	가림자 NS. /*PREDHA	가방 NS.
가격 NS.	가동 NS. /PRED2	가마 NS.	가변성 NS.
가결 NS. /PRED1/PRED3	가두 NS.	가마귀 NS. /ANM	가보 NS.
가경 NS.	가락 NS.	가마니 NS.	가부 NS.
가계 NS.	가락지 NS.	가멩 NS. /*PREDH	가부장 NS. /HUM
가공 NS. /PRED1/PRED3	가랑니 NS.	가멷 NS. /*PREDHA	가부좌 NS. /PRED2
가관 NS.	가랑이 NS.	가명 NS.	가불 NS. /PRED1/PRED3
가교 NS. /PRED2	가래 NS.	가모 NS.	가빈 NS. /HUM
가구 NS. /*PREDHA	가랭이 NS.	가묘 NS.	가사 NS.
가군 NS. /HUM	가로 NS.	가무 NS. /*PREDHA	가산 NS. /PRED1/PRED3
가금 NS. /ANM	가뢰 NS.	가문 NS.	가살 NS.
가난 NS. /*ADJH	가루 NS.	가물 NS.	가상 NS.; NS. /PRED1/PRED3
가내 NS.	가르마 NS. /NVS/*PREDHA	가물음 NS. /NVM	가설 NS. /PRED1/PRED3

< figure 2 - Extract of [DECOS-NS / V01] >

가공적이다 ADJS. /CM	가닥가닥하다 ADJS. /HM	가량스럽다 ADJS. /SM	가무댕댕하다 ADJS. /HM
가깝다 ADJS. /RM	가당찮다 ADJS. /RM	가련하다 ADJS. /HM	가무레하다 ADJS. /HM
가깝디가깝다 ADJS. /RM	가동적이다 ADJS. /CM	가렵다 ADJS. /RM	가무숙숙하다 ADJS. /HM
가깝하다 ADJS. /HM	가득가득하다 ADJS. /HM	가마노르께하다 ADJS. /HM	가무스럼하다 ADJS. /HM
가난하다 ADJS. /HM	가득하다 ADJS. /HM	가마득하다 ADJS. /HM	가무스레하다 ADJS. /HM
가날프다 ADJS. /RM	가들막가들막하다 ADJS. /HM	가마말쑥하다 ADJS. /HM	가무스름하다 ADJS. /HM
가느다랗다 ADJS. /RM	가들막하다 ADJS. /HM	가마무트름하다 ADJS. /HM	가무잡잡하다 ADJS. /HM
가느스레하다 ADJS. /HM	가뜩가뜩하다 ADJS. /HM	가마반드르하다 ADJS. /HM	가무족족하다 ADJS. /HM
가느스름하다 ADJS. /HM	가뜩하다 ADJS. /HM	가마반지르하다 ADJS. /HM	가무칙칙하다 ADJS. /HM
가늘다 ADJS. /RM	가뜬하다 ADJS. /HM	가맣다 ADJS. /RM	가무퇴퇴하다 ADJS. /HM
가늘디가늘다 ADJS. /RM	가랑가랑하다 ADJS. /HM	가무끄름하다 ADJS. /HM	가물가물하다 ADJS. /HM
가능하다 ADJS. /HM	가량맞다 ADJS. /MM	가무대대하다 ADJS. /HM	가뭇가뭇하다 ADJS. /HM

< figure 3 - Extract of [DECOS-ADJS / V01] >

Affixed forms and compound forms constitute other lexicons [DECOA / DECOC]. Given that some of the affixes (prefixes and suffixes) produce a considerable number of affixed forms, especially affixed **nouns**, we need complete lists of affixes in order to construct a lexicon of affixed items in a systematic way. The number of affixes taken into account in the current version is as followings <figure 4>:

Prefixes	*Suffixes*	*Pseudo-Nouns**	Total
950	900	180	2 030

< figure 4 - Numbers of *Pfx*, *Sfx*, and *PN* >

* Pseudo-Nouns are units that only occur in combinations with other nouns.

Notice that using a large corpus is indispensable for the construction of lexicons of affixed and compound items. However, remember that we can not obtain *automatically* the lists of these complex forms by combinating the lexicon of simple items with that of affixes, since there are too many homographs and therefore too many errors. In this case, we could try to establish syntactic or morphological rules that control wrong analyses and generations, but it seems to us that constructing valid general rules about all derivations and compositions would be a much more

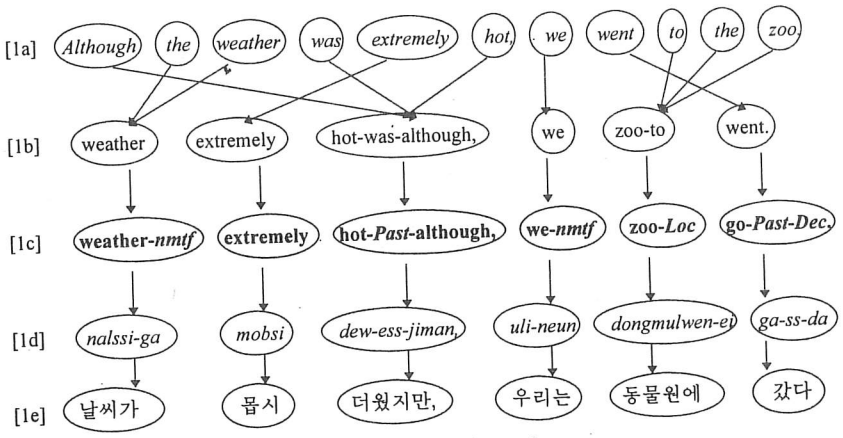
complicated task than describing all combined forms for each basic item. For the moment, the lexicons DECOA and DECOC are being constructed in such a procedure that their lists are estimated as much more complete than those we can find in existing dictionaries or in a large scale corpus: for example, the prefix '여 *ye-* [woman]' can be attached to nouns containing a semantic feature "*human*" such as 선생 *sensaing* [teacher], 간첩 *gancheb* [spy], 사장 *sajang* [boss]. We emphasize that, however, we do not search syntactic or semantic rules to list up these nouns, since all of the nouns with "*human*" feature do not admit the prefix '여 *ye-*': nouns denoting family relations or status such as 어머니 *emeni* [mother], 삼촌 *samchon* [uncle], 과부 *gwabu* [widow] do not accept this prefix: they already contain a gender marker; likewise, nouns describing human qualities such as 바보 *babo* [fool], 깍쟁이 *ggagjaingi* [miser] do not admit the prefix '여 *ye-*' either. Therefore, the list of nouns prefixed by '여 *ye-*' should be built up by examination of all simple nouns with "*human*" feature.

2. Lexicons of *N-Postpositions*, *A-Postpositions* and *V-Postpositions* DECO-POST

2.1. Strings delimited by separators

Let us recall that, in Korean, there are grammatical function markers such as *nominative*, *accusative*, *dative* or *locative* (we call them ***Noun*-Postpositions** [*PostN*] 명사 활용어), which are linked to nouns without any blanks. Thus, as we mentioned above, a sequence composed of two strings like *'in Paris'* corresponds to one string *'Paris-Locative'* in Korean. Likewise, verbs appear as conjugated forms like in English or French, but the inflectional suffix sets (we call them ***Verb*-Postpositions** [*PostV*] 동사 활용어) include several types of suffixes such as tense marker, modality marker, aspect marker, sentence or string type marker or politeness marker. Besides, the order and combinational constraints are extremely complex. Adjectives in Korean also should be followed by inflectional suffix sets (we call them ***Adjective*-Postpositions** [*PostA*] 형용사 활용어): suffixes indicate all grammatical functions of adjectives, whereas it is a copulative verb such as *'be'* or equivalent verbs in English, or such as *'être'* or equivalent verbs in French that takes the markers indicating grammatical functions of adjectival strings.

Thus, whereas the following sentence in English [1a] contains 9 strings separated by blanks, the corresponding sentence in Korean is composed of 6 strings as shown in [1e] <figure 5>:



< Figure 5 >

It is obvious that an automatic analyzer in Korean could not recognize canonical forms of nouns, verbs or adjectives without information about associable postposition types. (look at the phase [1d] in the graph above. Except the adverbial string '몹시 *extremely*', all strings are composed of a basic item and (a) grammatical suffix(es): '날씨 ***weather*** - 가 *Nominative Postposition*', '더우 ***hot*** - 었 *Past* - 지만 *Conjunctive Postposition*', '우리 ***we*** - 는 *Nominative Postposition*', '동물원 ***zoo*** - 에 *Locative Postposition*' and '가 ***go*** - 았 *Past* - 다 *Declarative Postposition*').

Therefore, a machine-readable dictionary (MRD) should provide information about all these strings. One could intend to represent a complete list of all conjugated forms in a MRD, given that finding out general rules that cover all cases is much more complicated than listing them out.

However, the number of sequences of postpositions for each basic item is considerable: a simple noun can be followed by around 1 500 different sequences of postpositions, since several postpositions can combine with one another (e.g. *Dative-Modality-Modality* such as '에게-만-이라도'); a verb and an adjective can be linked to around 6 000 types of postposition combinations. Hence, for a dictionary containing 35 000 basic items (*cf. Korean electronic dictionary of simple items* **DECOS**), we can observe around 100 million strings as shown in the following table <figure 6>: it will be too huge to be presented in the form of a list in a MRD.

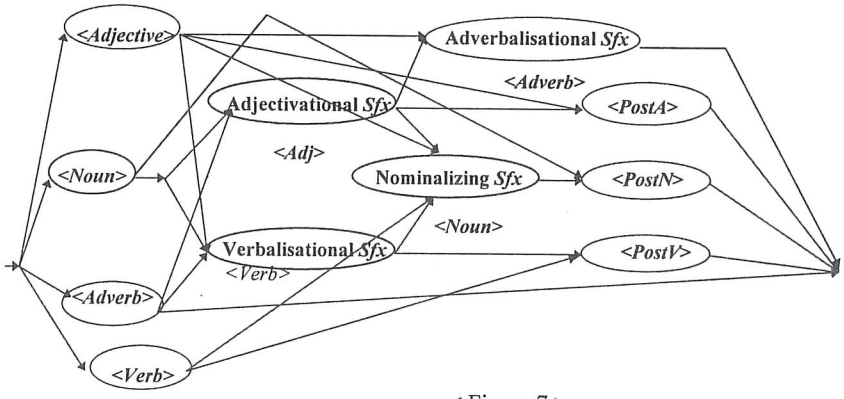
	Entry numbers in DECOS	Estimated String numbers
Simple Nouns	15 000	$15\,000 \times 1\,500 = 2.2 \times 10^7$
Simple Adjectives	5 300	$5\,300 \times 6\,000 = 3.2 \times 10^7$
Simple Verbs	7 500	$7\,500 \times 6\,000 = 4.5 \times 10^7$
Simple Adverbs / *Functional units*	7 000 200	7 000 * 200
Total	35 000	10^8

< Figure 6 >

(* We here do not take into account some *Adverb-Postpositions*, such as 도 *do*, 는 *neun*, 만 *man*, 만은 *man-eun*, 만이라도 *man-ilado*, etc: they are usually *modality markers*, and their number is limited to a few dozen.)

Remember that we here count only simple items (DECOS). If we also consider **affixed nouns** (such as *Prefix-Noun* or *Noun-Suffix* types), or **compound nouns** (such as *Noun-Noun* or *Adverb-Noun* types), given that the sizes of these lexicons are much larger than that of the lexicon of all simple items, it is clear that there is no point in building up a list of these strings.

Moreover, we observe very regularly nominalized forms of adjectives or verbs. These **nominalized forms** can be followed by *Noun-Postpositions* like any nominal sequence; so, for almost all adjectives and verbs, a considerable number of additional nominal strings can still be made up. The following graph represents a network of morphological relations among the main four lexical categories in Korean: *nouns*, *verbs, adjectives* and *adverbs* <figure 7>.



< Figure 7 >

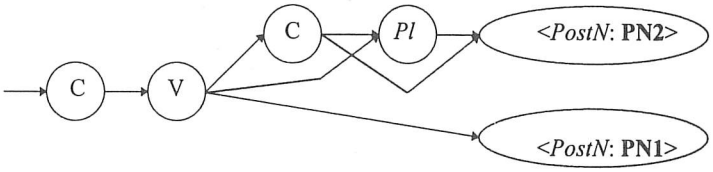
Therefore, for the time being, the method we have adopted in our lexical system DECO is to construct a lexicon of sequences of postpositions apart (DECO-*POST*), and to indicate morphological and inflectional information around each basic item, i.e. *nouns*, *adjectives* and *verbs*, in the lexicon DECOS: the sequences of postpositions are represented in the form of finite state automata (FSA).

2.2. Pinciples of sub-classification of postposition sets in DECO-POST

2.2.1. Classes of N-Postpositions

The combination of a *Noun* with a *PostN* is regular: the elements do not undergo morphological variations in connection. In the morphological view point, we can divide *PostN*s into two series: a series of *PostN*s that attaches to nouns with a **vocalic** ending; the other series of *PostN*s that attaches to nouns with a **consonantal** ending. We do not observe any morphological changes in the syllables in connection: neither in basic items themselves (nouns) nor in postposition sets.

Currently, *PostN* sets are sub-classified by morphological information types. The class **PN1** can be associated with **vowel ending** nouns, while the class **PN2** can follow **consonant ending** nouns. When a singular noun becomes a plural form (i.e. followed by plural marker '들 *deul* [pl]': it is the only grammatical marker of plurality in Korean), the postposition sets will be **PN2** type, since this marker ends in a consonant. We can represent these combinations as in the following graph <figure 8>:



< Figure 8 >

2.2.2. Classes of A-Postpositions and V-Postpositions

In the case of combinations of '*Adjectives* and *PostAs*' or '*Verbs* and *PostVs*', the syllables in connection undergo morphological variations. These variations can be summarized in three points:

A. Variations of the last syllable of basic items;
B. Variations of the first syllable of postposition sets;
C. Variations of both the last syllable of basic items and the first syllable of postposition sets.

The following examples illustrate these three cases respectively:

Ex-A. A verb '듣다 *deud(da)* [(to) hear]' becomes '들 *deul-*' on the left of postposition sets starting with a null consonant such as '어라 *-ela*' or '으면 *-eumyen*', whereas it remains unchanged '들 *deud-*' on the left of postposition sets starting with other consonants such as '고 *-go*' or '다가 *-daga*';

Ex-B. An adjective '착하다 *chagha(da)* [(to be) kind]' does not change when combined with postposition sets, but it requires a particular variant of the sequences of postpositions, when this sequence begins with a null consonant. Thus, postpositions starting with '여 *-ye*' such as '여서 *-yese*' or '였으므로 *-yesseumeulo*' only occur after verbs ending in '하 *ha*': '하여서 *ha-yese*', '하였으므로 *ha-yesseumeulo*';

Ex-C. A verb '굽다 *gub(da)* [(to) bake]' changes when the following postposition sets start with a silent consonant such as '으면 *-eumyen*' or '어 *-e*': the verbal string containing the first type of postposition will be '구우면 *gu-u-myen*', and the string with the second type will be a different type of fusion '구워서 *gu-we-se*'.

In the current version of the lexicon DECO-POST, we have integrated the morphological variants of postpositions [Ex-C type] and [Ex-B type] (the number of all postposition combination sets reaches to about 42000 in each of the cases of *PostA* and *PostV*). This dictionary provides information about the morphological types of basic items, i.e. adjectives and verbs. Here are samples of [DECO-POST / V01] <figure 9> and <figure 10>:

E \PN1 \PN2 *.nmtf .Acc .Postp*	같이까지가 \PN1 \PN2 *.Postp*	같이까지야 \PN1 \PN2 *.Postp*
가 \PN1 *.nmtf .Postp*	같이까지나 \PN1 \PN2 *.Postp*	같이까진 \PN1 \PN2 *.Postp*
가라도 \PN1 *.nmtf .Postp*	같이까지나마 \PN1 \PN2 *.Postp*	같이나 \PN1 \PN2 *.Postp*
가보다 \PN1 *.nmtf .Postp*	같이까지는 \PN1 \PN2 *.Postp*	같이나마 \PN1 \PN2 *.Postp*
가보다는 \PN1 *.nmtf .Postp*	같이까지라도 \PN1 \PN2 *.Postp*	같이는 \PN1 \PN2 *.Postp*
가보다도 \PN1 *.nmtf .Postp*	같이까지만 \PN1 \PN2 *.Postp*	같이는커녕 \PN1 \PN2 *.Postp*
가보단 \PN1 *.nmtf .Postp*	같이까지만도 \PN1 \PN2 *.Postp*	같이도 \PN1 \PN2 *.Postp*
같이 \PN1 \PN2 *.Postp*	같이까지만이 \PN1 \PN2 *.Postp*	같이라도 \PN1 \PN2 *.Postp*
같이가 \PN1 \PN2 *.Postp*	같이까지만이라도 \PN1 \PN2 *.Postp*	같이를 \PN1 \PN2 *.Postp*

< Figure 9 - Extract of [DECO-POST / V01] - ***N-Postpositions*** >

E(아) \PA1 *.TmDec .Conj*	ㄴ가는 \PA1 \PA2 \PA5 \PA7 *.CoDis*	ㄴ가만은 \PA1 \PA2 \PA5 \PA7 *.Conj*
E(아)? \PA1 *.TmInt*	ㄴ가는 \PA1 \PA2 \PA5 \PA7 *.Conj*	ㄴ가만이 \PA1 \PA2 \PA5 \PA7 *.Conj*
ㄴ \PA1 \PA2 \PA5 \PA7 *.Dtm*	ㄴ가도 \PA1 \PA2 \PA5 \PA7 *.Conj*	ㄴ가만이라도 \PA1 \PA2 \PA5 \PA7 *.Conj*
ㄴ가 \PA1 \PA2 \PA5 \PA7 *.Conj*	ㄴ가라도 \PA1 \PA2 \PA5 \PA7 *.Conj*	ㄴ가뿐만아니라 \PA1 \PA2 \PA5 \PA7 *.CoDis*
ㄴ가? \PA1 \PA2 \PA5 \PA7 *.TmInt*	ㄴ가마저 \PA1 \PA2 \PA5 \PA7 *.Conj*	ㄴ가뿐만아니라 \PA1 \PA2 \PA5 \PA7 *.Conj*
ㄴ가가 \PA1 \PA2 \PA5 \PA7 *.Conj*	ㄴ가마저도 \PA1 \PA2 \PA5 \PA7 *.Conj*	ㄴ가뿐아니라 \PA1 \PA2 \PA5 \PA7 *.CoDis*
ㄴ가나 \PA1 \PA2 \PA5 \PA7 *.Conj*	ㄴ가만 \PA1 \PA2 \PA5 \PA7 *.Conj*	ㄴ가뿐아니라 \PA1 \PA2 \PA5 \PA7 *.Conj*
ㄴ가나마 \PA1 \PA2 \PA5 \PA7 *.Conj*	ㄴ가만도 \PA1 \PA2 \PA5 \PA7 *.Conj*	ㄴ가야 \PA1 \PA2 \PA5 \PA7 *.Conj*

< Figure 10 - Extract of [DECO-POST / V01] - ***A-Postpositions*** >

2.3. Final syllable types of adjectives and verbs

As we mentioned above, we have 5 300 adjectives and 7 500 verbs in the current version of our lexicon DECOS. We have classified them according to the final syllable types: we have obtained 151 types for adjectives and 325 types for verbs. These classes can be regrouped according to the required postposition types. The following tables represent respectively the first entries of these types: <figure 11> and <figure 12>.

Type	A-1	감	Type	A-11	곧	Type	A-21	길	Type	A-31	낫
Type	A-2	갑	Type	A-12	곱	Type	A-22	깊	Type	A-32	낮
Type	A-3	갈	Type	A-13	꿸	Type	A-23	깜	Type	A-33	넓
Type	A-4	갛	Type	A-14	교	Type	A-24	깝	Type	A-34	넘
Type	A-5	걸	Type	A-15	굳	Type	A-25	껌	Type	A-35	녹
Type	A-6	검	Type	A-16	굵	Type	A-26	껍	Type	A-36	높
Type	A-7	겁	Type	A-17	굽	Type	A-27	꼽	Type	A-37	눅
Type	A-8	겋	Type	A-18	궂	Type	A-28	나	Type	A-38	늘
Type	A-9	결	Type	A-19	글	Type	A-29	낡	Type	A-39	늙
Type	A-10	겹	Type	A-20	기	Type	A-30	납	Type	A-40	늦

< Figure 11- Final syllable types of *Adjectives* >

Type	V-1	가	Type	V-11	겪	Type	V-21	굳	Type	V-31	기
Type	V-2	갈	Type	V-12	견	Type	V-22	굴	Type	V-32	긷
Type	V-3	갉	Type	V-13	결	Type	V-23	굶	Type	V-33	길
Type	V-4	감	Type	V-14	고	Type	V-24	굽	Type	V-34	깁
Type	V-5	갖	Type	V-15	골	Type	V-25	귀	Type	V-35	까
Type	V-6	갚	Type	V-16	곪	Type	V-26	그	Type	V-36	깎
Type	V-7	개	Type	V-17	곯	Type	V-27	글	Type	V-37	깔
Type	V-8	걷	Type	V-18	곱	Type	V-28	긁	Type	V-38	깜
Type	V-9	걸	Type	V-19	괴	Type	V-29	금	Type	V-39	깨
Type	V-10	게	Type	V-20	구	Type	V-30	긋	Type	V-40	꺾

< Figure 12 - Final syllable types of *Verbs* >

III. Perspectives

So far, we have discussed the method we have adopted in constructing Korean electronic lexical system DECO. Given that existing editorial dictionaries are hardly reliable, the use of a large Korean corpus should be required. For the construction of the lexicon of **simple items** [DECOS / V01], we have consulted existing editorial dictionaries, but formal and explicit principles have been used: accurate attribution of parts of speech is done; morphological and syntactic information is indicated in a coherent way.

A large corpus is required especially when we build up lexicons of **affixed items** and **compound ones**. However, even if the scale of the corpus is considerably extended, on one hand, we can never avoid lack of lexical items; on the other hand, we cannot only expect appropriate identification of affixed and compound forms. Therefore, the exhaustive description of combinations for a given item will be indispensable to obtain all correct sets related to each item and only them: powerful general grammars that cover all cases do not exist.

The current version of our lexical system DECO provides, on one hand, all simple items, classified by parts of speech: *nouns* (NS), *adjectives* (ADJS), *verbs* (VS), *adverbs* (ADVS), *functional units* (FUS); and the affixes: *prefixes* (PF), *suffixes* (SF), *pseudo-nouns* (PN). On the

other hand, it contains a lexicon named DECO-POST that provides all postposition sets, i.e. *Noun-Postpositions* (PostN), *Adjective-Postpositions* (PostA), *Verb-Postpositions* (PostV): these sets are represented in the form of finite state automata.

The lexicons of affixed items and compound ones should be developed in an exhaustive and coherent way, by using combinatorial procedures based upon the lexicons of simple items and that of affixes. Besides, all information about the conjunction of lexical items (DECOS) and grammatical items (DECO-POST) has to be described in detail.

Reference

Clemenceau, David, 1993, *Structuration du lexique et reconnaissance de mots dérivés*, PhD thesis, Paris: Univ. Paris7.

Courtois, Blandine, 1987, *Dictionnaire éléctronique du Laboratoire d'Automatique Documentaire et Linguisique pour les mots simples du français* (DELAS), Rapport Technique of LADL, n°17, Paris: Univ. Paris7.

Courtois, Blandine, 1989, DELAF: *Dictionnaire éléctronique du LADL pour les mots fléchis du français*, Rapport Technique of LADL, N°20, Paris: Univ. Paris7.

Grand Dictionnaire Encyclopédique Larousse, 1982, Paris: Larousse.

Grand Robert de la Langue Française, 1986, Paris: Le Robert.

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

Gross, Maurice, 1989, La construction de dictionnaires élétroniques, *Annales des Télécommunications,* tome 44 N°1:2, Issy-les-Moulineaux / Lannion:CNET.

I, Hi-Seung, 1988, *Guge Dai Sajen* (Korean Dictionary), Seoul: Minjungselim.

Kim, Myung-Cheol; Seo Kwang-Jun; Jun Kyung-Heon, 1992, *Development of Natural Language Interface*, Report in ETRI, Korea.

Nam, Jee-Sun, 1991, *Etablissement du corpus des adjectifs coréens*: Rapport technique N° 30, Paris: Institut Blaise Pascal, University Paris 7.

Nam, Jee-Sun, 1992, Corpus des adjectifs coréens: Constitution et classification, *XVème Congrès International des Linguistes*, Québec: University Laval.

Nam, Jee-Sun, 1994, Représentation de la combinatoire des variantes consonantiques et vocaliques et de la combinatoire des suffixes de conjugaison des adjectifs en coréen, *Papers in Computational Lexicography Complex '94*, ed. by Ferenc Kiefer, Gabor Kiss et Julia Pajzs, Budapest: Linguistics Institute, Hungarian Academy of Sciences.

Nam, Jee-Sun, 1994, *Dictionnaire des noms simples du coreen*, Rapport technique N°46, LADL.

Nam, Jee-Sun, 1996, *Dictionary of Korean simple verbs : DECOS-VS / V01,* Rapport technique N° 49, Laboratoire d'Automatique Documentaire et Linguistique, Université Paris 7.

Nam, Jee-Sun, 1996, *Lexicon of Korean predicative terms classified by morphological ending forms, Vol. II. : Predicates except* Hada *ending ones*, IGM Rapport N° 96-9, Institut Gaspard Monge, Université de Marne-la-Vallée.

Nam, Jee-Sun, 1996, *Dictionary of N-Postpositions, A-Postpositions and V-Postpositions in Korean :DECO-POST / V01*, Rapport technique N° 51, LADL, Université Paris 7.

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

Roche, Emmanuel, 1993, *Analyse syntaxique transformationnelle du français par transducteurs et lexique-grammaire,* PhD thesis, Univ. Paris 7.

Silberztein, Max, 1993, *Dictionnaires éléctroniques et analyse automatique de textes - le systeme Intex,* Paris: Masson.

Sin, Gi-Chel; Sin, Yong-Chel, 1990, *Sai Ulimal Keun Sajen* (New Korean Dictionary), Seoul: Samseng Chulpansa.