Knowledge-Rich Approach to Automatic Grammatical Information Acquisition:

Enriching Chinese Sketch Engine with a Lexical Grammar

Chu-Ren Huang
Inst. Of Linguistics
Academia Sinica
churen@sinica.edu.tw

Wei-Yun Ma
Inst. Of Info. Science
Academia Sinica
ma@iis.sinica.edu.tw

Yi-Ching Wu
Inst. Of Linguistics
Academia Sinica
d898702@oz.nthu.edu.tw

Chih-Ming Chiu
Inst. Of Linguistics
Academia Sinica
henning@hp.iis.sinica.edu.tw

Abstract

This paper discusses the implementation of a knowledge-rich approach to automatic acquisition of grammatical information. Our study is based on Word Sketch Engine (Kilgarriff and Tudgell 2002). The original claims of WSE are two folded: that linguistic generalizations can be automatically extracted from a corpus with simple collocation information provided that the corpus is large enough; and that such a methodology is easily adaptable for a new language. Our work on Chinese Sketch Engine attests to the claim the WSE is adaptable for a new language. More critically, we show that the quality of grammatical information provided has a directly bearing on the result of grammatical information acquisition. We show that when provided with a knowledge rich lexical grammar, both the quantity and quality of the extracted knowledge improves substantially over the results with simple PS rules.

1 Background: Word Sketch Engine and Automatic Acquisition of Grammatical Information

The original goal of corpus-based studies was to provide 'a body of evidence' for more theoretical linguistic studies (Francis and Kucera 1965). However, corpus-based studies evolved with the improvements made in electronic data manipulation, making of automatic acquisition of grammatical information a goal of computational linguistics, computational lexicography, as well as theoretical corpus linguistics. Previous works that made significant contribution to the study of automatic extraction of grammatical relation includes Sinclair's (1987) work on KWIC, Church and Hanks' (1989) introduction of Mutual Information, and Lin's (1998) introduction of relevance measurement.

Kilgarriff and colleagues' work on Word Sktech Engine (WSE) makes a bold step forwards in automatic linguistic knowledge acquisition (Kilgarriff and Tudgell 2002, Kilgarriff et al. 2004). The main claim is that a 'gargantuan' corpus[1] contains enough distributional information about most grammatical dependencies in a language such that the set of simple collocational patterns will allow automatic extraction of grammatical relations and other grammatical information. Crucially, the validity of the extracted information does not rely on the preciseness of the rules or the perfect grammaticality of the data. Instead, WSE allows the presence of ungrammatical examples in the corpus and the possibility for collocational patterns to occasionally identify the wrong lexical pairs. WSE assumes that these anomalies will be statistically insignificant, especially when there are enough examples instantiating the intended grammatical information. In addition, WSE relies on Salience measurement to rank the significance of all attested relations. Salience is calculated by MI of a relation multiplied with the frequency of the relation, in order to correct MI's bias towards low frequency items. WSE follows Lin's (1998) formulation of MI of relations, where $\|w_1, R, w_2\|$ stands for the frequency of the relation R between w_1 and, w_2. A wild card * can occurs in place of w_1, R, or w_2 to represent the all cases. Hence MI between w_1, and w_2 given a relation R is given below (Kilgarriff and Tudgell 2002):

---

[1] The required corpus size was not specified in WSE literature. However, we estimate from existing work that for WSE to be efficient, corpus scale must be 100 millions words or above.

$$(1). \ I(w1, R, w2) = \log(\frac{\| *, R, * \| \ x \ \| w1, R, w2 \|}{\| w1, R, * \| \ x \ \| *, R, w2 \|})$$

With Salience ranking, WSE gives a one page summary of the most significant grammatical behaviors of any given word. The report includes SUBJ, OBJ, modifier, coordination, etc. WSE is also able to calculate Sketch differences between two sketches, and create automatic thesauri that underline the comparisons between the synonym pairs based on sketch similarity.

2 Previous Work: The Preliminary Implementation of Chinese Word Sketch

A crucial claim of the WSE is that this methodology can be easily adapted to new languages. That is, each language would require a different set of collocational patterns for relation extraction. WSE has been successfully ported to Czech and Irish (Kilgarriff et al. 2004). And work has done to produce a prototype of Chinese Sketch Engine (called CSE I hereafter for easy reference, Kilgarriff et al. 2005).

One issue not addressed in previous literature on WSE or similar work on automatic extraction of grammatical information is how much can existent grammatical knowledge help. While WSE requires only simple collocational information, it was not clear if more sophisticated grammatical information will help or hurt the result of the WSE. Three previous adaptation of the WSE, including Kilgarriff et al.'s (2005) adaptation of CSE I, replies heavily on transferring the original BNC-based templates to a different language and achieved reasonable results. However, there have been observations that they seem to miss some language-specific grammatical behaviors.

Word Sketch uses regular expressions over POS-tags to formalize rules of collocation patterns. CSE I utilizes 11 collocating patterns to extract all grammatical relations and only one pattern for the simplest verb-object relation as shown as (2)

(2) Collocating Pattern for Object from CSE I

1:"V[BCJ]" "Di"? "N[abc]"? "DE"? "N[abc]"? 2: "Na" [tag!= "Na"]

("XXX" represents XXX is a regular expression, "XXX"? represents XXX appears zero or one time, "XXX"{a,b} represents XXX appears a~b times.)

In (2), the 1: and 2: identify the two collocated components. Between the components, zero or one particle may appear (denoted by "Di"?), zero or one processor may appears (denoted by any_noun? "DE"?), and zero or one noun-modifier may appears (denoted by "N[abc]"?)

Huang et al. (2005) pointed out that the prototype version of CSE I did not deal with the prevalent non-canonical word orders in Chinese (3). In addition, we also noticed that it fails to identify grammatical relations when an argument lies some distance away from a verb because of internal modification (4). Chinese objects often occur in pre-verbal positions in various pre-posing constructions, such as topicalization.

(3)

 a. 全穀麵包，**吃**了很健康。
    *quan.gu mian.bao, **chi** le hen jian.kang*
    whole-grain bread, **eat** LE very healthy
    'Eating whole-grain bread is very healthy.'

 b. 有人嘗試要**將**這荷花**分類**，卻越分越累。
    *you ren chang.shi yao **jiang** zhe he.hua **fen.lei**, que yue fen yue lei*
    someone try to **JIANG** the lotus **classify**, but more classify more tired
    'People have tried to decide what category the lotus belongs in, but have found the effort taxing.'

(4). 他　只　**吃**了　一　口　　**飯** …
　　　T*a zhi* **chi** *let yi kou* **fan**
　　　s/he only **eat** ASP one mouthful **rice**

　　Such examples led to the question of whether the simple collocation rules adapted in Kilgarriff et al. (2005) was sufficient and if a knowledge-rich approach would yield better results.

3　Porting ICG lexical grammar as collocation patterns

3.1　Motivating a knowledge-rich approach

The important design criteria of WSE is that salience statistics is compiled based on relational tuples such as $\{w_1, R, w_2\}$. This is a crucial decision since word-based lexical statistics itself does not offer enough grammatical information, while it is hard to obtain enough information-rich parsed trees for statistic studies. It is interesting to observe that Kilgarriff et al. (2002) obtained only 70 million tuples (types) based on the 100 million words BNC. In terms of elements that need to be traced, this is indeed comparable to a general bi-gram model and definitely less complex than models that allows any lexical bi-gram without adjacency conditions. The reason for the reduction in complexity is because the collocational patterns serve as filters that disregard non-significant relations. Based on this model, a set of collocational patterns that contains richer grammatical information will enable the sketch engine to better identify grammatical relation tuples and render more precise grammatical information. Ideally, the most effective collocational patterns are those with explicit annotations of the targeted grammatical relations. Hence we propose to port a lexical grammar with argument annotation as WSE collocational patterns.

3.2　Introducing ICG

The Information-based Case Grammar (Chen and Huang 1992) is a unification-based formalism proposed specifically for Chinese language processing. ICG is a head-driven lexical grammar in the sense that all grammatical information is encoded on the verb. Each verb is encoded with a set of basic patterns (BP) which stipulate the possible structural instantiations of that verb as well as the positions of participant roles (called Case) for each verb. There are over 100 templates of patterns corresponding to each verb sub-class. In the Academia Sinica CKIP lexicon, over 40,000 verbs are annotated with ICG information. Each verb starts with a default assignment according to its verbal sub-class, with the template information manually corrected based on corpus data and linguistic analysis. Obviously, not unlike the Levin classes for English (Levin, 1993), each BP is repeated and shared by a number of verb sub-classes. Both the BP information and the Verb sub-classes information will be utilized in our adaptation of Chinese Sketch Engine (referred to as CSE II hereafter).

(5). ICG Basic Patterns for Stative Pseudo-transitive Verb (VI)

　　　EXPERIENCER<GOAL[PP[對]]<VI
　　　EXPERIENCER<VI<<GOAL[PP[於]]
　　　THEME<GOAL[PP{對、以}]<VI
　　　THEME<VI<<GOAL[PP[於]]
　　　THEME<VI<<SOURCE[PP{自、於}]
　　　THEME< SOURCE[PP{歸、爲}]<VI
　(A<B represents B appears behind A. A<<B represents B appears immediately behind A)

3.3　Implementation: Preparation of Corpus and Grammar

There are two steps in the implementation of CSE II: the first step is corpus preparation, and the second is grammar adaptation. For corpus, we follow CSE I and use the LDC Chinese Gigaword Corpus because of its size (over 11 billion characters) and its coverage of both traditional and simplified characters. The Gigaword Corpus was fully automatically segmented and tagged using the Academia Sinica tagset and tagtool (Ma and Huang 2006). Our work in adaptation for CSE II includes resolution of categorical ambiguity for nominalization and improvement of unknown word resolution.

For grammar adaptation, we concentrate on exploiting lexically encoded ICG grammatical knowledge. Since the corpus was tagged with Academia Sinica tagset, the verb-subclass information for each verb is specified. Hence we can utilize the structural information from ICG BP. Since the tagged corpus has identified the verb-subclasses, we are able to correctly identify different grammatical relations, even though two verbs may share the same local structure. For instance, many verbs share the [..PP V NP] structure. However, for pseudo-transitive verbs (VB), contrary to naïve structural assignment, it is the object of PP that has the Object role for the matrix sentences, as illustrated in 6. Such structural mismatches are easily resolved when the sub-class tag information is unified with ICG BP information.

(6). 醫生幫病人開了三次刀
   [yisheng]np [bang bingren]pp [kai]
   v le[san ci da]np
   Doctor for patient operate ASP three CLS operation
  'The doctor operated on the patient three times.'

A further crucial step that we take in grammar adaptation is to allow a dependency relation that is separated by several constituents. Recall that a crucial motivation for the design of WSE is because parsing would be too time- and labor- consuming and would not yield highly reliable results. However, without parsing, it would be difficult to identify a head of a complex object, or a preposed object. Based on ICG grammar, we observe that such behaviors are often dependent on the verb sub-classes and can be captured. An illustrating example is the identification of a preposed object of a pseudo-transitive verb (VB).

(7)
 a. 村莊(object) 明天將 被 夷為平地(VB11)
   cunzhuang mingtian jiang bei yiweipingdi
    village tomorrow will BY level-to-the-ground
   'The village will be leveled to the ground tomorrow.'

 b. begin time1 location time1 adv? passive_prep adv_string 1:"V[BCJ].*" [tag!="DE"]

Note that the rule allows CSE II to ignore the temporal NP closer to the verb and pick the initial NP as the Object (denoted by 'location' in (7b), which is a noun phrase describing a location. Complete definition is given in appendix). Another set of rules utilizes the fact that successive Chinese nouns of a NP are head final. Hence, in order to determine which noun of a NP is the head Object, we stipulate that it has to be the Object which must not precede another noun. A relation and the rule that it accounts for are given in (8). Note that the NP stands for a noun-head following zero, one, or two noun-modifiers. The rule correctly pick jingguan 'sight-and-view' and not the noun-modifier preceding it (i.e. gongyuan 'park') as the Object of pohuai 'to damage'.

(8)
 a. 大量 的 遊客 破壞(VC2) 公園 景觀(object)
  daliang de youke pohuai   gongyuan jingguan
  large-number DE tourists damage park sight-and-view
 'Large number of tourists damaged the sight-and-view of the park.'

 b. 1:"VC.*" (particle|prep)? NP not_noun

   (NP is defined as "…noun_modifier{0,2} 2:noun…". Complete definition is given in appendix.)

The 32 definitions and 80 collocating patterns are designed for all Chinese grammatical relations according to their sub-classes. Note that the English grammar has 39 definitions but only 40 collocating patterns. We can safely say the CSE II grammar contains richer structural information. Of

the 80 patterns, 20 of them are for verb-object relations. The complete list is given in Appendix I for reference. Please note that the number of rules are greater than the number (11 collocating patterns in all, one of them for verb-object relation) for CSE I. The new grammar took the Word Sketch Engine over 7 hours to compile. But once complied, the composition of word sketch for each word could be done in real time. Please note that, based on the compile log, each of the 20 object rule are useful and applied to at least 2, 7515415,56153,713,[103] times. This clearly shows that all new rules are basic and necessary.

4 Results and Evaluation

At the time of submission, only spot-checks of the results have been performed. Overall evaluation is still being conducted and results will be available in the final paper. The spot-checking so far does show clear and evident improvements over CSE I.

(9) Object Recall Comparison

	CSE I	CSE II
hong2 (red)	0	0
pao3 (run)	0	8,704
kan4 (look)	32,350	64,096
da3 (hit)	26,016	47,182
song4 (give)	0	76,378
shuo1 (say)	0	20,350
xiang1xin4 (believe)	0	52,373
quan4 (persuade)	0	3,852

The recall data comparing in (9) underlines the drastic improvement of CSE II over CSE I. For simple transitive verbs (the state verb *kan4* and the activity verb *da3*), CSE II recall almost twice as many objects as CSE I. For more complex verb (ditransitive *song4*, as well as all types of clause taking verbs *xiang1xin4, xiang1xin4,* and *quan4*), CSE I fails to identify any of their objects, while CSE II. does correctly extract their objects. On the other hand, for intransitive verbs, CSE I and CSE II both correctly extract no object relations for the state verb *hong2*. The fact that CSE II extracted some object relations for the activity verb *pao3*, although with relatively low frequency, is worth noting. Upon further examination, we found that many of the objects extracted have habitual readings, such as pao2 *ma3la1song1'* runs marathon' or idiomatic reading *pao3 bai2tie3* '(of a politician) runs from one funeral to another'. These are additional senses of the lemma *pao3* that to take objects. In sum, the recall comparison data shows improvement of both quality and quantity.

In order to contrast the quality of the extracted grammatical knowledge, we take the verb *chi1* 'to eat' for a more in-depth analysis. For *chi1,* only 23,421 objects were identified by CSE I, while we identified 33,038 objects with the richer grammar patterns in CSE II. This is an improvement of over 42% in terms of recall and a substantial quantitative gain. In terms of quality improvement, we observed that the following three objects are among the top 20 collocates identified by CSE II, but no by CSE I.

(10). a. 飯 fan4 rice 802 70.96 (4),
    b. 虧 kui disadvantage 329 59.24 (12)
    c. 苦頭 ku3tou2 suffering 194 58.71 (14)

Note that the three numbers following each object is its frequency (as object of *chi*), its saliency in this relation, and its saliency ranking (in parentheses). Note that both *chi-gui* 'to be taken advantage of' and *chi-kutou* 'to suffer' are both idiom chunks, and expected to be among the most salient collocating objects of chi. However, since they both allow frequent internal modification (e.g. *chi zhangsan de an kui,* 'been taken advantage of in the dark by Zhangsan'), a simple collocation pattern

such as adopted by CSE I fails to identify them. Our adaptation in CSE II took internal modification into consideration and successfully identified them. The case with *fan* is even more general and potentially more interesting in terms of extracting basic collocation. Rice is undoubtedly the most typical conceptual object of *chi* 'to eat' and it occurs frequently in the corpus. However, CSE I only identified 266 instances of *fan* as object of *chi,* even less than the 427 instances of *binglang* 'beetlenut'. This is because *fan* represents a basic and generic concept and is rarely used along without modification. Since it often does not occur in concatenation with the verb, the simple collocation pattern of CSE I cannot identify it. We can see in (10) that CSE II identifies 802 instances of *fan* as object of *chi*, a recall improvement of over 200%. In addition, CSE II shows that *fan* as object of *chi* is almost twice as frequent as *binglang* (450). This fact is more consistent with our knowledge of the Chinese language and a clear indication that our adaptation successfully corrected the biased introduced by the incomplete grammatical knowledge of in CSE I.

Nevertheless, a recall of instance *fan* as an object improves over 200% in terms of its identification., misplace of instance *fan* as a subject still remain. As CSE II shown, 718 instances of *fan* as a subject require us to modify our grammar adaptation. In fact, instance *fan* will never serve as a grammatical relation of subject, hence collocation patterns of *object/object_of* ought to be adapted according to its sub-classes. In view of 718 instances of *fan* as a subject, we found that both of *mei* and *you* that precede a POS "Na" play a significant role in marking a object and identifying topicalization.

(11) 保證 災民 **有** 飯 吃、**有** 衣 穿、有 住處。
　　baozheng　zaimin you fan chi　、yao yi chuan　、yao zhuchu
　　ensure　victims　YOU rice eat　、YOU clothes wear　、have　dwelling place
　　'We ensure that the victims will have rice to eat, clothes to wear and have dwelling places.'

(12) 他 相信 水利處 工作 人員 不會 **沒有** 飯 吃。
　　ta　xiang xin shuilichu gongzuo renyuan buhui meiyou fan chi
　　he　believe　department of irrigation and engineering　staff　won't MEI rice eat
　　'He believes that the staff in department of irrigation and engineering will have rice to eat.'

The examples above reveal that an object is likely to be identified between *mei / you* and "VC.*". In that case, collocating pattern for object in CSE II can be altered and added to extract the very collocation of verb_object like this,

[word="沒"|word="沒有"|word="有"]NP adv_string 1:"VC.*" [tag!="DE"]

Although this collocating pattern cannot capture all the topicalized objects (e.g. 我飯吃完就走了。), it seems to help identify instance *fan* as an object as illustrated in CSE II, or rather, it helps to mark the object in another collocation of verb_object indeed. In addition to the collocating pattern illustrated above, there exists a sentence pattern that helps to point out the topicalized objects,

(13) 他 經常 是 一頭 扎進 實驗室 就 **連** 飯 **都** 顧不上 **吃** 。
　　ta jingchang shi yitou zhajin　shi yan shi jiu lian fan dou　gubushang chi
　　often　SHI completely invest laboratory jiu LIAN rice DOU unconcernedly eat
　　'He often invest such much time in the laboratory that he forgets to eat.'

In example (15), it represents a predication of *lian-dou* pattern and the topicalized object *fan* is in-between. Therefore, we may extract the collocation of verb_object stated as below,

[word="連"] NP [word="都"| adv_string] 1:"VC.*" [tag!="DE"]

Hereby, we still are confronted with one problems as below, though *lian-dou* construction seems to help extract all the topicalized objects:

(14) 這種 飯 就 連 乞丐 都 不 吃。

    zhezhong fan jiu lian qigai dou bu chi

    This sort rice jiu LIAN beggar DOU not eat

    'Even a beggar won't eat this sort of rice. '

    In the light of the sentence (14), we are certain to come up with more refined grammar adaptation to capture the real topicalized object that instantiates in the natural language realization. Identifying an object to be a topicalization is really a thorny problem in terms of grammatical knowledge; even though the above suggested collocating patterns advance the identification of object as a sub-class, the goal is aimed to extract all sorts of topicalized objects in CSE II.

5   Conclusion

In this paper, we showed that rich grammatical knowledge can be utilized to improve results of automatic grammatical information acquisition. In particular, we applied the ICG lexical grammar to the WSE to support the claim that its methodology can be easily adapted in different languages. We also introduced richer grammatical information to the finite state specifications of WSE, such as verb subclasses and displaced arguments. We showed that a knowledge-rich approach substantially improves the quality and quantity of extracted grammatical information with our adaptation of verb sub-class and argument Basic Pattern information from ICG. We believe that this is an encouraging result for future developments of automatic grammatical acquisition research since acquired grammatical knowledge can provide feedback to the system in order to improve future results. In addition, our work also points to a potentially productive way for theoretical and computational linguists to collaborate.

References

Chen, Keh-jiann and Chu-Ren Huang. 1990. Information-based Case Grammar. Proceedings of the 13th International Conference on Computational Linguistics (COLING '90).   Vol.ii.54-59. Helsinki, Finland. August 20-25th.

Church, Ken. W. and Hanks, Patrick. 1989. Word association norms, mutual information and lexicography. Proceedings of the 27th Annual Meeting of ACL. Pp. 76-83. Vancouver.

Huang, Chu-Ren, Adam Kilgarriff, Yiching Wu, Chih-Min Chiu, Simon Smith, Pavel Rychly, Ming-Hong Bai, and Keh-jiann Chen. 2005. Chinese Sketch Engine and the Extraction of Collocations. Proceedings of t the Fourth SigHan Workshop on Chinese Language Processing. Jeju, Korea.

Kilgarriff, Adam and Tugwell, David. Sketching Words. 2002. In Marie-Hélène Corréard (ed.): Lexicography and Natural Language Processing. A Festschrift in Honour of B.T.S. Atkins. Euralex.

Kilgarriff, Adam, Pavel Rychlý, Pavel Smrz and David Tugwell. 2004. The Sketch Engine. Proceedings of EURALEX, Lorient, France.

Kilgarriff, Adam, Chu-Ren Huang, Pavel Rychlý, Simon Smith, and David Tugwell. 2005. Chinese Word Sketches. ASIALEX 2005: Words in Asian Cultural Context. Singapore.

Levin, Beth. 1993. English Verb Classes and Alternations: A Preliminary Investigation. University of Chicago Press.

Lin, Dekang. 1998. Automatic retrieval; and clustering of similar words. Proceedings of COLING-ACL. Montreal. 768-774.

Ma, Wei-yun and Chu-Ren Huang. 2006. Uniform and Effective Tagging of a Heterogeneous Giga-word Corpus. Proceedings of LREC 2006. Genoa.

Sinclair, John. M. (editor). 1987. Looking Up: an account of the COBUILD project in lexical computing. Collins.

Appendix: Complete Direct Object Patterns of CSE II

Definition of Symbols

define(`adj',`("A"|"VH11"|"VH13"|"VH21"|"V.*" "DE")')
define(`dm',`("Neqa"|"Neu"|"Neqa" "Nf.*"|"Neu" "Nf.*"|"Nf.*")')
define(`dm1',`("Neqa"|"Neqa" "Nf.*"|"Neu" "Nf.*"|"Nf.*")')
define(`adj_string',`dm? adj{0,2}')

define(`prep',`"P.?.?"')
define(`rel_clause',`("P.?.?" [tag!=".*Y"]{1,5}|[tag!=".*Y"]{1,3}) ("Ng"|[word="的"] [word="時候"])')

define(`particle',`"Di"|"T"|"I"')
define(`passive_prep',`[word="被"|word="給"|word="遭"|word="由"|word="挨"|word="受"|word="備受"|word="遭受"]')

define(`exclude_prep',`[word!="被" & word !="給" & word!= "遭" & word!="由" & word!="挨" & word!="受" & word!="備受" & word!="遭受" & word!="把" & word!="將" & word!="向"]')

define(`adv',`([tag="D.*" & tag!="DE"]|"V.*" "DE")')
define(`adv_string',`adv{0,2}')

define(`noun',`[tag="N[abcdhf].*" & tag!="Nbc.*" & tag!="Ncd.*" & word!="者" & word!="們"]')
define(`not_noun',`[tag="N[abcdhef].*"|tag="Nbc.*"|tag="Ncd.*"]')
define(`not_nounandDE',`([tag="N[abcdhef].*" & tag!="DE"]|[tag="Nbc.*"]|[tag="Ncd.*"])')

define(`noun_without_NcNd',`[tag="N[abh].*" & tag!="Nbc.*" & word!="者" & word!="們" & word!="年" & word!="月" & word!="日"]')
define(`noun_modifier ',`[tag="N[abcd].*" & tag!="Ncd"]')

define(`time',`adv_string adj_string "Nd.*"{0,5} 2:"Nd" "Ng"?')
define(`location',`adv_string adj_string "Nc.*"{0,5} 2:"Nc" "Ncd"? "Ng"?')
define(`time1',`(prep? adv_string adj_string "Nd.*"{1,6} "Ng"?)?')
define(`location1',`(prep? adv_string adj_string "Nc.*"{1,6} "Ncd"? "Ng"?)?')

define(`NP',`([tag!=".*Y" & tag!="DE"]{1,3} "DE")? adv_string adj_string noun_modifier{0,2} 2:noun "Ncd"?')
define(`NP_without_NcNd',`([tag!=".*Y" & tag!="DE"]{1,3} "DE")? adv_string adj_string noun_modifier{0,2} 2:noun_without_NcNd')
define(`NP1',`([tag!=".*Y" & tag!="DE"]{1,3} "DE")? adv_string adj_string noun_modifier{0,2} noun "Ncd"?')
define(`NP1_without_NcNd',`([tag!=".*Y" & tag!="DE"]{1,3} "DE")? adv_string adj_string noun_modifier{0,2} noun_without_NcNd')

define(`begin',`("COMMACATEGORY"|"PERIODCATEGORY"|"VE.*"|"VK.*")')
define(`end',`("COMMACATEGORY"|"PERIODCATEGORY"|"Ca.*"|"Cb.*")')

Collocating patterns of Object/Object_of

*DUAL
=Object/Object_of

    1:"V[ACFJKL].*" (particle|prep)? NP not_noun

    1:[tag="VH12"|tag="VH14"|tag="VH16"|tag="VH17"|tag="VH22"] (particle|prep)? NP not_noun

    [word="把"|word="將"|word="向"] NP adv_string 1:"VB.*" [tag!="DE"]

    [word="把"|word="將"] NP adv_string 1:"VC.*" [tag!="DE"]

    NP_without_NcNd time1 location1 time1 adv? passive_prep adv_string 1:"V[BCJ].*" [tag!="DE"]

    NP_without_NcNd time1 location1 time1 adv? passive_prep NP1 adv_string 1:"V[BCJ].*" [tag!="DE"]

    begin time1 location time1 adv? passive_prep adv_string 1:"V[BCJ].*" [tag!="DE"]

    begin time1 location time1 adv? passive_prep NP1 adv_string 1:"V[BCJ].*" [tag!="DE"]

    1:"V[BD].*" (particle|prep|[word="給"])? NP not_noun

    1:"VG.*" [word="爲"|word="作"]? NP not_noun

    [word="對"|word="以"] NP adv_string 1:"VI.*" [tag!="DE"]

    1:"VI.*" [word="自"|word="於"]? NP not_noun

    [word="對"] NP adv_string 1:"VJ.*" [tag!="DE"]

    1:"VD.*" (particle|prep|[word="給"])? NP1 NP not_noun

    [word="把"|word="將"|word="向"] NP adv_string 1:"VD.*" (particle|"Ng"|"Ncd.*")? end

    NP_without_NcNd time1 location1 time1 adv? passive_prep adv_string 1:"V[DE].*" (particle|"Ng"|"Ncd.*")? end

    NP_without_NcNd time1 location1 time1 adv? passive_prep NP1 adv_string 1:"V[DE].*" (particle|"Ng"|"Ncd.*")? end

    1:"VE.*" (particle|prep)? NP (particle|"Ng"|"Ncd.*")? end

    1:"VE.*" (particle|prep)? NP1 NP (particle|"Ng"|"Ncd.*")? end

    [word="向"] NP adv_string 1:"VE.*" (particle|"Ng"|"Ncd.*")? End