Korean–English MT and S-TAG

Mark Dras and Chung-hye Han
Macquarie University *Simon Fraser University*

1. Introduction

An early motivation for Synchronous TAG (S-TAG) (Shieber and Schabes, 1990) was machine translation (Abeillé, Schabes and Joshi, 1990). Abeillé *et al* note that traditionally difficult problems outlined by Dorr (1994)—for example, categorial, thematic, conflational, structural and lexical divergences—have been used to argue for the necessity of an explicit semantic representation. However, many of these divergences are not problems for an S-TAG-based approach. Synchronous TAG translation models thus allow us to explore the question of the extent to which a semantic representation is actually necessary.

S-TAG was redefined by Shieber (1994) for both theoretical and practical reasons, introducing the requirement that the derivation trees of target and source be isomorphic. Under this definition it has been noted (Shieber, 1994; Dras and Bleam, 2000) that there are mappings that cannot be described under S-TAG. This was the motivation for meta-level grammars (Dras, 1999), by which two TAG grammars can be paired while retaining their original properties, as under standard S-TAG, allowing for a description of mappings that include unbounded non-isomorphisms (Dras and Bleam, 2000).

This work on exploring how S-TAG (with and without meta-level grammars) can be used for MT has only been applied to languages that are closely related—English, French, Italian and Spanish. In this paper we aim to take a much more widely differing pair of languages, English and Korean, to investigate the extent to which syntactic mappings are satisfactory.

English and Korean have a wide range of differences: rigid SVO word order in English vs verb-final with free word order in Korean, the largely analytic structure of English vs the agglutinative structure of Korean with its complex morphology, optional subject and object and the absence of number and articles in Korean, and many others. These all suggest that a meta-level grammar will be necessary as there are various many-to-one or many-to-many mappings between derivation tree nodes (i.e., there will be few cases where a single elementary tree corresponds to another single elementary tree, which has been the case with closely related languages).

Although there is an implemented Korean/English MT system that includes a TAG Korean parser as a source language analysis component (Han *et al.*, 2000), this system as a whole is based on Meaning Text Theory (Mel'čuk, 1988), an enriched dependency formalism. Thus, it requires a conversion component that converts the TAG parser output to a dependency notation. As pointed out in Palmer *et al.* (2002), however, this conversion process resulted in a loss of crucial information such as predicate-argument structure encoded in TAG elementary trees, which had negative consequences in the translation results. This then provides further motivation to explore the feasibility of applying a single TAG-based formalism to modeling and implementing a Korean/English MT system.

As a first step towards exploring the extent to which an S-TAG style approach can successfully model these widely different languages, we have taken from a parallel English-Korean Treebank twenty examples of divergent constructions (see Appendix). Each half has roughly 50,000 word tokens and 5,000 sentences. While the annotation guidelines for the Korean half was developed in Han, Han and Ko (2001) for this corpus, the English half follows the guidelines already developed for Penn English Treebank (Bies *et al.*, 1995), as closely as possible. The example pairs represent structures including copula, predicative/attributive adjective, passive, causative, interrogative, relative clause, complex verb, and modal construction, among others. We find that using a TAG-based meta-level grammar to model Korean/English correspondences for machine translation is quite feasible.

2. Analyses

In this section we discuss two example pairs of sentences, taken from the parallel Treebank, that illustrates several divergences, and how an S-TAG with meta-level grammar can handle them. The trees we use for the subgrammars for the sentences are extracted automatically from the Treebank using Lextract (Xia, Palmer and

 Proceedings of the Sixth International Workshop on Tree Adjoining Grammar and Related Frameworks (TAG+6), pp. 206–215. Universitá di Venezia.

Joshi, 2000).[1]

2.1. Korean complex NP vs. English modal

The sentence pair in (1) represents a modal construction. The key divergence is that the Korean uses a noun complement structure, while the English uses a modal adjective structure:

(1) 전차들은 그 능력을 개활지에서 충분히 발휘할 수 있습니다.
tank-Plu-Top that ability-Acc open-terrain-Loc fully show-Adnominal possibility be-Past-Decl
Tanks are able to fully demonstrate their potential in open terrain.

A closer but less natural translation of the Korean is *The possibility that tanks fully demonstrate their potential in open terrain exists*; the noun representing *possibility* is modified by an adnominal clause. The corresponding English translation contains *be able to* followed by an infinitival clause. The derivation trees are as in Figure 1, and the Lextract elementary trees grouped according to the translation pairing in Figure 2.

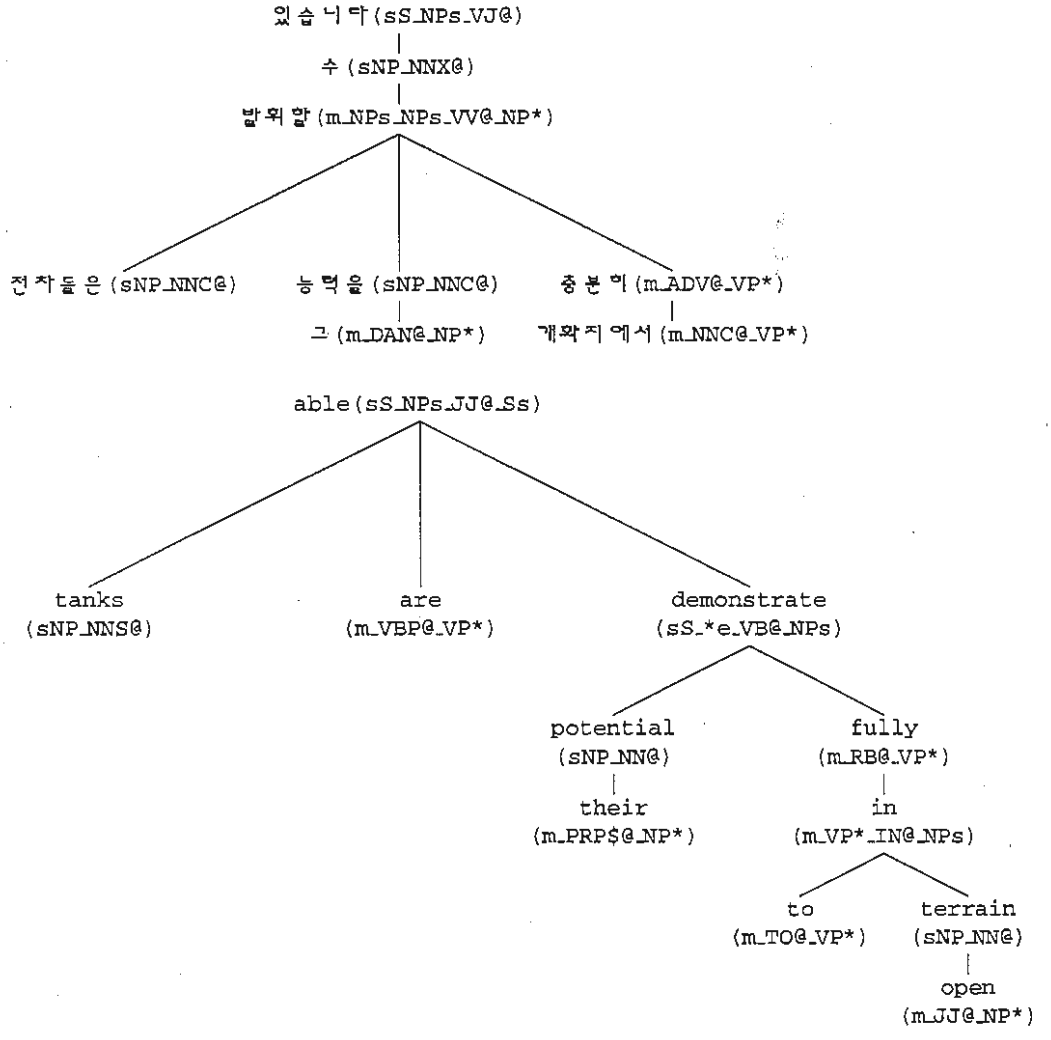

Figure 1: Derivation trees for (1)

1. Note that the Lextract trees do not contain features, although the corresponding Korean XTAG (Han *et al.*, 2000) trees do. We will make use of the features where necessary.

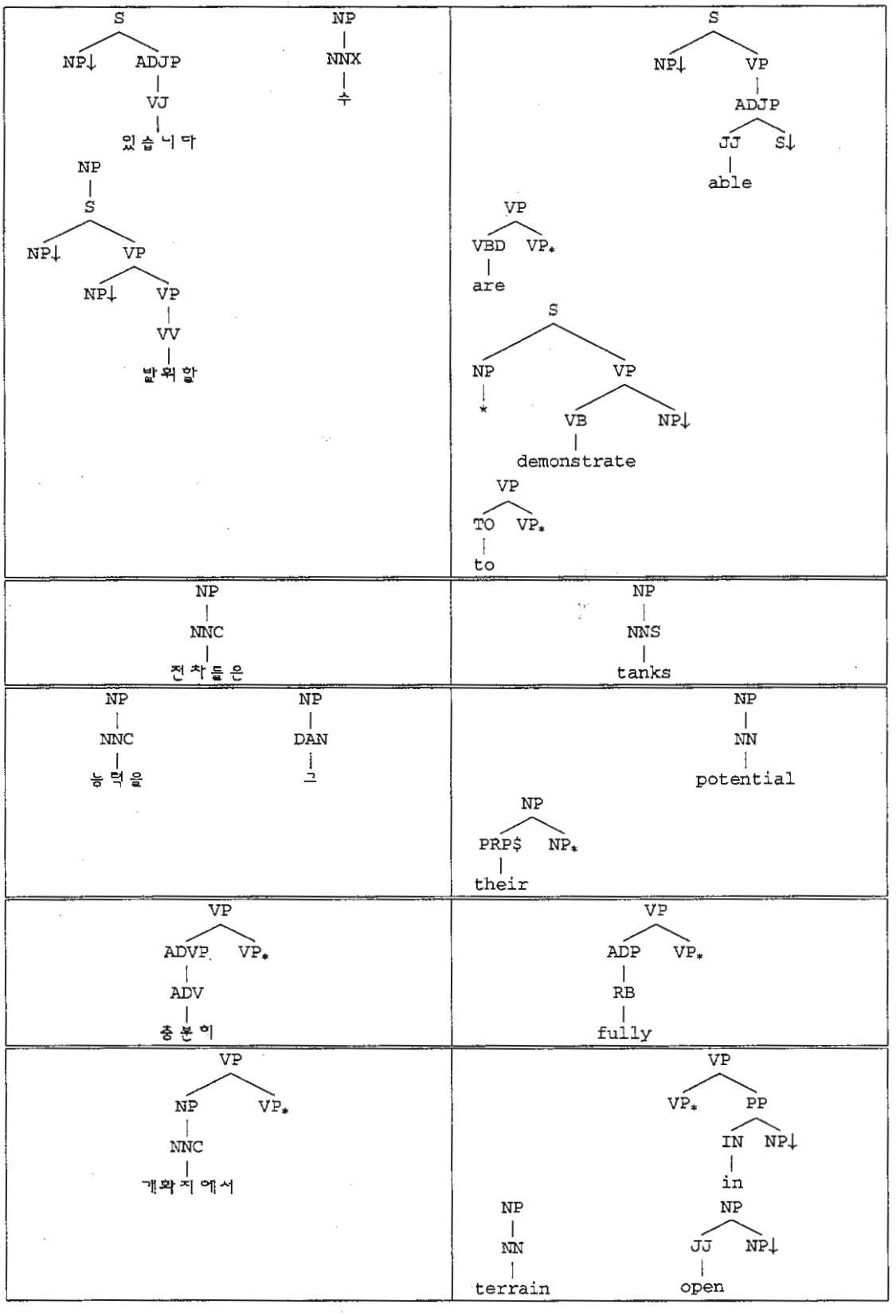

Figure 2: Lextract elementary trees for (1)

The trees are clearly far from isomorphic. The relationship between *able* and *are* is inverted between the corresponding Korean 있 슬 니 다 (*be*) and 수 (*possibility*), although *demonstrate* is the child of *able* (발 휘 할 and 수 respectively) in both. Most crucially, however, the infinitival *to* in English, attached to *demonstrate*, has no corresponding element in Korean; rather, *to* and *demonstrate* correspond to the single 발 휘 할 in Korean. But, given TAG's approach to modification, an unbounded number of modifiers (*fully*, the PP headed by *in*) can be inserted between *demonstrate* and *to*, giving an unbounded non-isomorphism. In other examples we have noted that this unbounded non-isomorphism is quite prevalent, occurring inter alia with nouns and determiners.

Other divergences attested in (1) are that *tanks* is an argument of *able*, but 전 차 들 은 (*tanks*) is an argument of 발 휘 할 (*demonstrate*); and that the preposition *in* is represented by the suffix 서, a type of correspondence that occurs frequently because of the analytic–agglutinative language mismatch. Using the algorithm of Dras (1999), however, it is possible to construct a meta-level grammar to characterize appropriate paired substructures in the trees, as in Figure 3. The basic principle is that the divergent material is captured by the multi-level tree pairs (such as 19–A), in particular in cases with unbounded non-isomorphisms, where the recursive material (such as 19–D and 19–E) is factored out. The other structures that are not a cause of the isomorphism violation continue to to be paired by single-level tree pairs (either as in 19–B, or in cases not illustrated here where there is a single node corresponding to a lexicalized tree plus a substitution node).[2]

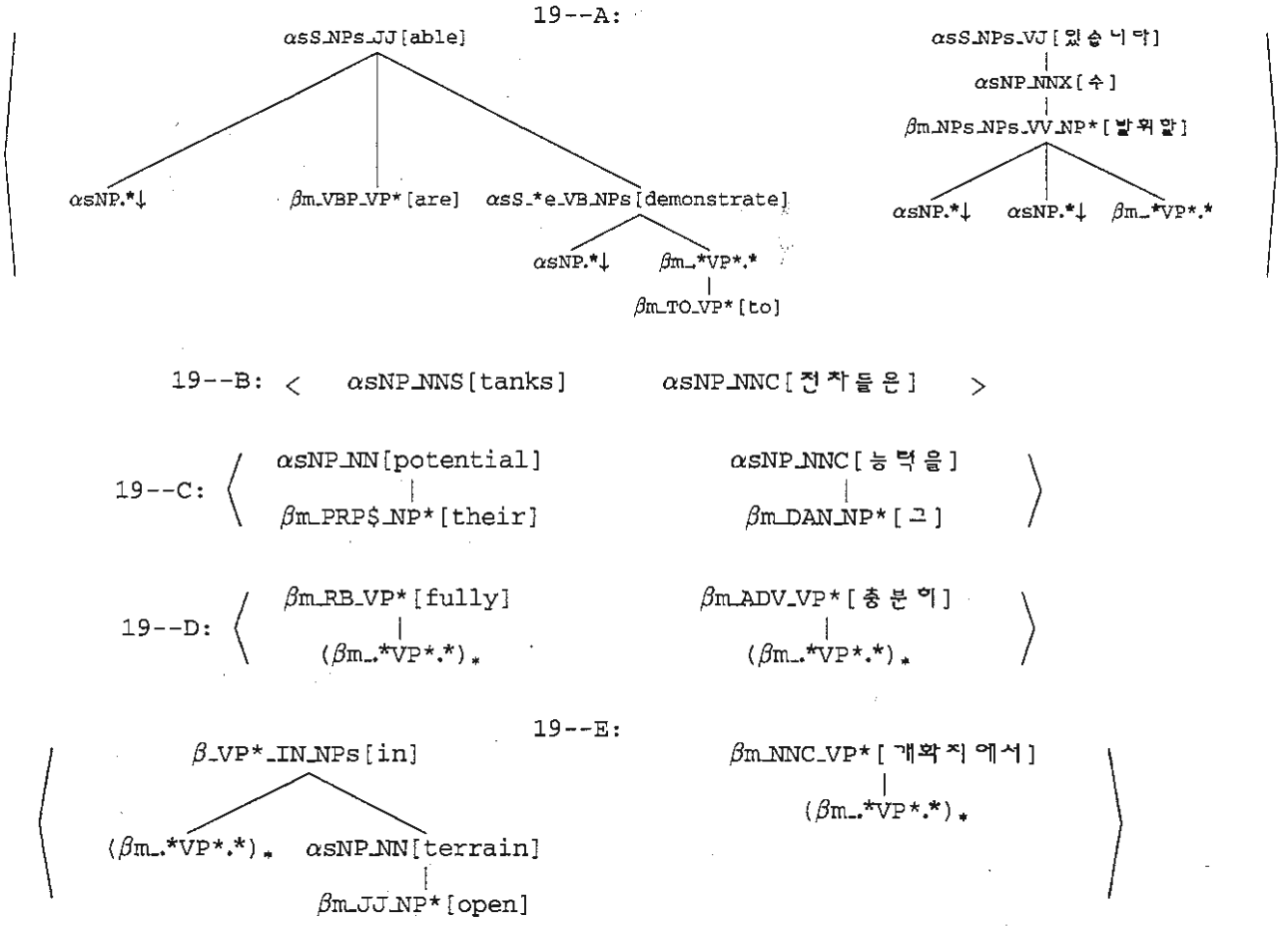

Figure 3: Meta-level grammar for (1)

The groupings that arise from the algorithm are fairly intuitive. 19–A represents the concept *the ability of X to demonstrate Y* (X here being *t*anks and Y *p*otential), with two consequent argument slots, and one slot where a modifier can be adjoined marked βm_*VP*.*. [3] 19–B and 19–D are straightforward; 19–C aggregates the nodes because in general Korean does not use determiners, so an English noun and determiner correspond to a

2. If a pairing of isomorphic trees was expressed by a meta-level TAG, all trees would be single-level.
3. This regular expression represents a node where any tree with a root whose label matches can adjoin; technically this is

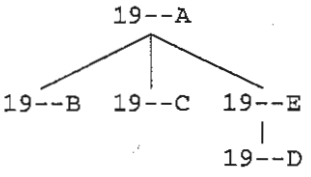

Figure 4: Meta-level derivation for (1)

single unit in Korean (although this is not the case here, we follow that general principle); and 19–E represents the correspondence between the English PP *in open terrain* and the single Korean 개활지 에서. Under this meta-level grammar we have isomorphic meta-level derivation trees for English and Korean with structure as in Figure 4.

Note that, as a next step, the obvious generalisation is to have a single parametrized tree pair in cases like 19–A and 19–E. From 19–A we will have the same structure for *X are able to demonstrate Y*, *X are able to see Y*, and so on, with a Korean correspondent for each choice of verb. From 19–E we will have the same structure for *in open terrain*, *near open terrain*, and so on, with a corresponding Korean suffix for each choice of preposition. With the suffixes in Korean XTAG represented by features, the approach would be similar to that of Abeillé, Schabes and Joshi (1990) for cases where the French and English share a feature-related attribute like number.

For the example here it could be argued that perhaps *to* 'should' be in the same tree as *demonstrate*, and that in general there should not be separate elementary trees for function words. Frank (2001) argues for functional elements to be part of lexical elementary trees, and this is the principle used in building the large-scale French TAG grammar, although each has different ideas as to which trees functional elements should be included in. However, part of the aim of translating with S-TAG is to use already existing grammars; there are not special separate grammars for translation that have matching choices about function word treatment. And it is unlikely that all choices would match in any case, for example with determiners, which would be likely separate in English and French, but not in Korean.[4]

2.2. Copula constructions

Korean does not have an explicit copula; this gives rise divergences as in the sentence pair (2).

(2) 경기관총 분대장은 중사입니다.
light-machinegun squad-leader-Top sergeant-Cop-Decl
The light machinegun squad leader is a sergeant.

This is not problematic because of the way in which TAG conventionally represents copular constructions, where the predication is the root of the derivation and the copula is adjoined in. Derivation trees are as in Figure 5.

The feature of interest in this translation is the absence of Korean determiners, as mentioned in the previous example. The combined noun-determiner in English thus corresponds to only the noun in Korean; and there can be recursive intervening material (such as *light*, *machinegun* and *squad* between *the* and *leader*). Thus we again have an unbounded non-isomorphism, and we handle it with a meta-level grammar as in Figure 6.

3. Discussion

In our analysis of twenty sentence pair types (see Appendix) chosen to illustrate particular divergences not typically found between closely related languages, a TAG meta-level grammar is basically adequate for describing the mapping between them, using the algorithm of Dras (1999).

because the labels are really just features (Kasper *et al.*, 1995; Dras, Chiang and Schuler, 2002). Thus, slightly confusingly, there are three types of asterisk in a meta-level grammar. Firstly, there is the asterisk that is part of the name of an XTAG or Lextract tree; this is indicated by a normal aterisk *. Secondly, there is the asterisk to indicate a regular expression over these names; this is indicated by a bold asterisk **. Thirdly, there is the asterisk to indicate a footnode in a meta-level auxiliary tree; this is indicated by a subscripted asterisk $*$. All three occur in, for example, the right projection of 19–E.

4. In fact, the fact that F-TAG includes function words in lexical trees and the XTAG English grammar does not suggests that a meta-level grammar may be useful there also.

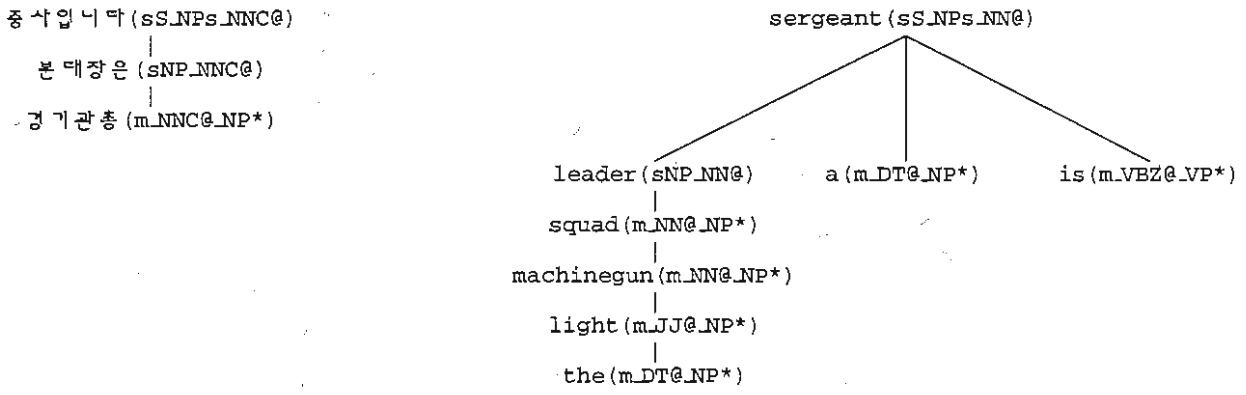

Figure 5: Derivation trees for (2)

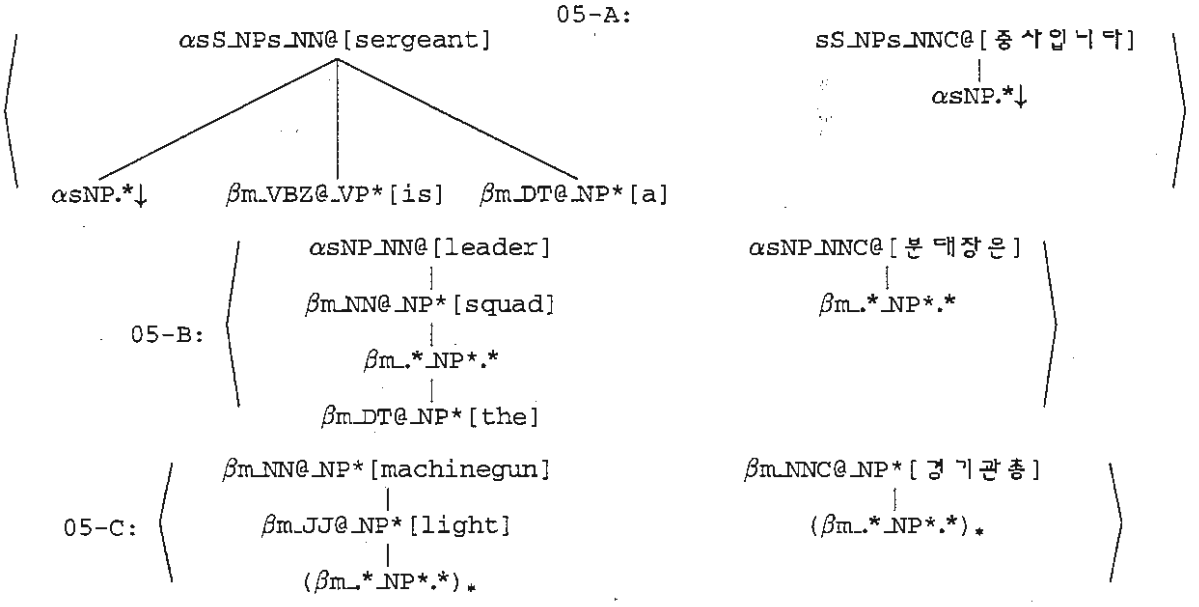

Figure 6: Meta-level grammar for (2)

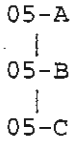

Figure 7: Meta-level derivation for (2)

The major exception is with some adverbial modifiers that can occur both sentence-initially and adjacent to VP without any semantic difference. Because TAG is fundamentally a constituent-based formalism, it is necessary to have two different trees for such modifiers (e.g., *soon*) depending on the location of the modifier (S-rooted and VP-rooted). Thus, in a sentence pair as in (3) in which *now* is VP-adjoined and 지 금 ('now') is S-adjoined, it is not possible to build a reasonable TAG meta-level grammar. To see this examine the derivation trees given in Figure 8. Most nodes pair up straightforwardly (*on schedule* pairing with 계획 대로, with the Korean containing a suffix to parallel *on*); the exceptions are the nodes for *now* and *proceeding*, which would have to be grouped together because of the different dominance relations (계획 대로 being immediately dominated by 진 행되 고, but there being the possibility of unbounded intervening material between *proceeding* and *now*). This grouping of *proceeding* and *now* would be fairly unprincipled, as *now* is a case of recursive material that does not belong in an elementary tree pair at the meta-level. That is, a meta-level grammar is still formally adequate, but linguistically undesirable.

(3) 지 금 그 공 격 준 비 가 계획 대로 진 행되 고 있 습 니 다
now that attack preparations-Nom plan-as proceed-Pass-Auxconn be-Past-Decl
The attack preparations are now proceeding on schedule.

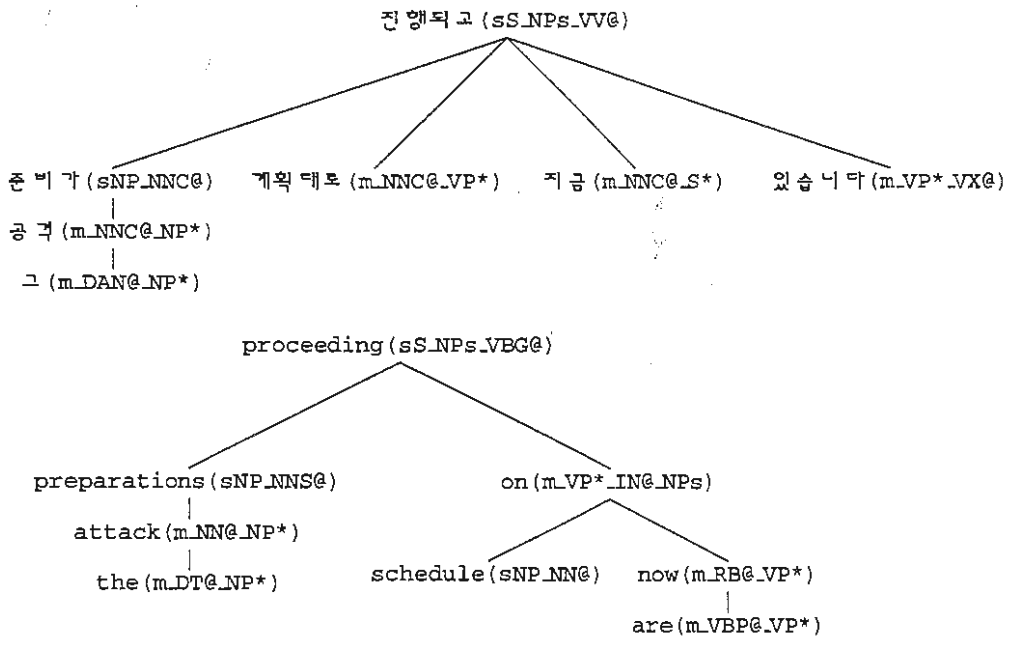

Figure 8: Derivation trees for (3)

However, no semantic difference will result if *now* were sentence-intial in the English, or if 지 금 ('now') were adjacent to the verb 진 행되 고 ('proceed') in the Korean. This means that even if the Treebank translation does not allow a meta-level grammar, one is possible just by moving the modifier. From our initial exploration, then, a meta-level grammar appears to be a promising candidate for describing English-Korean translation.

The next stage of the work is to build a prototype system and use a Lextract-like approach to extract a meta-level grammar from the parallel Treebank. Lextract already provides us with elementary and derivation trees for Treebank pairs; the algorithm of Dras (1999) gives a systematic method for identifying paired substructures in derivation trees. Further, our prototype system will include a generation component (for Korean and/or English, depending on what the target language is) that generates derivation and derived trees from a given meta-level derivation structure.

A. Sample divergences

#simple declaratives

(4) 제가 대대장한테 관측 사항을 보고하였습니다.
I-Nom battalion-commander-to observation stuff-Acc report-Past-Decl
I reported my observations to the battalion commander.

#declarative with object scrambling

(5) 그들의 규모나 명칭은 저는 모릅니다.
their size-or designation-Top I-Top don't-know-Decl
I don't know their sizes or designations.

#attributive adjective

(6) 도로의 상태와 적 정황 중요한 요소가 됩니다.
road-Gen condition-Conj enemy situation be-important-Adnominal factor-Nom be-Decl
Road conditions and the enemy situation are key factors.

#predicative adjective

(7) 대대 정치 군관의 권한은 대단히 크지요.
battalion political office-Gen authority-Top very extensive-Decl
The authority of the battalion political officer is very extensive.

#copula sentence

(8) 경기관총 분대장은 중사입니다.
light-machinegun squad-leader-Top sergeant-Cop-Decl
The light machinegun squad leader is a sergeant.

#Korean passive morphology → English passive form

(9) 부대 명칭은 통상 암호로 하달됩니다.
unit designation-Top normally code-in transmit-Pass-Decl
Unit designations are normally transmitted in code.

#Korean passive morphology → English active form

(10) 지금 그 공격 준비가 계획대로 진행되고 있습니다.
now that attack preparations-Nom plan-as proceed-Pass-Auxconn be-Decl
The attack preparations are now proceeding on schedule.

#Korean active form → English passive form

(11) 그러니까 더 이상 그런 선전에는 속지 마!
so any more that propaganda-by-Top deceived don't
So don't be deceived by that propaganda anymore!

#lexical causative

(12) 눈보라가 교통을 마비시키었다.
snowstorm-Nom traffic-Acc paralysis-Cause-Past-Decl
The snowstorm paralyzed the traffic.

#structural causative

(13) 중대 특무장은 중대 성원들이 무기와 탄약을 갖도록
compay first-sergeant-Top company members-Nom weapons-Conj ammunition-Acc have
합니다.
make-Decl
The company first sergeant ensures that the members of the company have the weapons and ammunition.

#structural causative

(14) 그러면 대대부가 대대 공급반한테 탄약을 수송하도록 하었습니다
then battalion-HQ-Nom battalion supply-section-to ammunition-Acc transport-Caus do-Past-Decl
Our battalion HQ then had the ammunition brought in by the battalion's supply section.

#yes-no question

(15) 소대장의 호출대호가 바뀌었는가?
platoon-leader-Gen call-sign-Nom changed-Past-Int
Has the call sign of the platoon leader been changed?

#wh-question

(16) 전파 지향성 공중선은 어떤 무전기를 사용하는가?
radio(wave) directional antenna-Top what radio-Acc use-Int
What types of radios is the inclined beam antenna used with?

#relative clause

(17) 그 무선 전화수가 사용한 책은 컸다.
that radiotelephone operator-Nom use-Adnom book-Top big-Past-Decl
The book that the radiotelephone operator used was big.

#Korean morpheme → English word

(18) 포병 지원 부대가 대대에 배속되면 이들도 초단파망을
artillery support unit-Nom battalion-To attach-Pass-when these-persons-also microwaves-net-Acc
사용합니다.
use-Decl
When artillery support units are attached to the battalion, they would use the VHF network also.

#Korean noun and light verb → English verb

(19) 송신기나 수신기는 가끔 손질을 해야 합니다.
transmitter-and receiver-Top occasionally cleaning-Acc do-Auxconn must-Decl
One must clean the transmitters and receivers occasionally.

#Korean complex verb → English verb and adverb

(20) 그 편지 저한테 돌티어 주십시오.
that letter me-To handover-Auxconn give-Imp
Please give me back the letter.

#Korean complex verb → English verb and preposition

(21) 분대장은 그 부상병의 눈을 조심스럽게 들여다 보았습니다.
squad-leader-Top that wounded-soldier-Gen eye-Acc carefully take-in look-Past-Decl
The squad leader carefully looked into the eyes of the wounded soldier.

#Korean complex noun phrase → English modal auxiliary verb construction

(22) 전차들은 그 능력을 개활지에서 충분히 발휘할 수 있습니다.
tank-Plu-Top that ability-Acc open-terrain-Loc fully show-Adnominal possibility be-Past-Decl
Tanks are able to fully demonstrate their potential in open terrain.

#Korean intransitive verb → English transitive

(23) T-54 탱크는 발연했읍니다.
T-54 tank-Top smoke-emit-Past-Decl
The T-54 tank emitted smoke.

References

Abeillé, Anne, Yves Schabes and Aravind K. Joshi. 1990. Using lexicalized tree adjoining grammars for machine translation. In *Proceedings of the 13th International Conference on Computational Linguistics (COLING '90)*, Helsinki, Finland, August.

Bies, Ann, Mark Ferguson, Karen Katz and Robert MacIntyre. 1995. Bracketing Guidelines for Treebank II Style Penn Treebank Project.

Dorr, Bonnie. 1994. Machine Translation Divergences: A Formal Description and Proposed Solution. *Computational Linguistics*, 20(4):597–633.

Dras, Mark. 1999. A meta-level grammar: redefining synchronous TAG for translation and paraphrase. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL '99)*, pages 80–87.

Dras, Mark and Tonia Bleam. 2000. How Problematic are Clitics for S-TAG Translation? In *Proceedings of the Fifth International Workshop on Tree Adjoining Grammars and Related Formalisms (TAG+5)*, pages 241–244, Paris, France.

Dras, Mark, David Chiang and William Schuler. 2002. On Relations of Constituency and Dependency Grammars. *Journal of Language and Computation*.

Frank, Robert. 2001. Phrase Structure Composition and Syntactic Dependencies. MS. Johns Hopkins University, June.

Han, Chung-hye, Benoit Lavoie, Martha Palmer, Owen Rambow, Richard Kittredge, Tanya Korelsky, Nari Kim and Myunghee Kim. 2000. Handling Structural Divergences and Recovering Dropped Arguments in a Korean/English Machine Translation System. In John S. White, editor, *Envisioning Machine Translation in the Information Future*, Lecture Notes in Artificial Intelligence. Springer-Verlag, pages 40–53. Proceedings of the Association for Machine Translation in the Americas, AMTA 2000.

Han, Chunghye, Na-Rae Han and Eon-Suk Ko. 2001. Bracketing Guidelines for Penn Korean Treebank. Technical Report IRCS-01-10, Institute for Research in Cognitive Science, University of Pennsylvania.

Han, Chunghye, Juntae Yoon, Nari Kim and Martha Palmer. 2000. A Feature-Based Lexicalized Tree Adjoining Grammar for Korean. Technical Report IRCS-00-04, Institute for Research in Cognitive Science, University of Pennsylvania.

Kasper, Robert, Bernd Kiefer, Klaus Netter and K. Vijay-Shanker. 1995. Compilation of HPSG to TAG. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics (ACL '95)*.

Mel'čuk, Igor. 1988. *Dependency syntax: theory and practice* . Albany: State University of NY Press.

Palmer, Martha, Chung-hye Han, Anoop Sarkar and Ann Bies. 2002. Integrating Korean analysis components in a modular Korean/English machine translation system. MS. University of Pennsylvania and Simon Fraser University.

Shieber, Stuart M. 1994. Restricting the weak-generative capability of synchronous tree adjoining grammars. *Computational Intelligence*, 10(4).

Shieber, Stuart M. and Yves Schabes. 1990. Synchronous tree adjoining grammars. In *Proceedings of the 13th International Conference on Computational Linguistics (COLING '90)*, Helsinki, Finland, August.

Xia, Fei, Martha Palmer and Aravind Joshi. 2000. A Uniform Method of Grammar Extraction and Its Applications. In *Proceedings of EMNLP 2000*.