Processing Korean Dialogue: a first attempt ...

Martin Weisser
Erlangen University
martin.weisser@rzmail.uni-erlangen.de

1. Introduction

This paper represents a first attempt at devising a method to analyse Korean (transactional) dialogues automatically. As such, it should not be regarded as a fully developed methodology yet, but rather seen as an 'outsider's' view on how best to approach such an analysis. The approach envisaged is an adaptation of the one developed during the compilation of the SPAAC corpus[1], where approximately 1,250 English dialogues from two domains where successfully annotated using an annotation tool written in Perl/Tk (see Weisser, '02a & 04a).

In the present paper, I will attempt to demonstrate that this approach should be relatively easily transferable to Korean. In order to do this, I will point out similarities, as well as differences, between the steps necessary for analysing English and Korean.

2. The Data

The data used to develop my preliminary ideas is based on dialogues from the domain of hotel bookings and was provided to me by Prof. Choe, Jae-Woong of the Linguistics Department, Korea University. The initial work was conducted on a very small-scale sample of only 10 dialogues, although the corpus comprises 88 dialogues in total.

3. Taxonomy of Generic Units

The SPAAC approach I am planning to port to Korean is mainly based on the idea that there are certain types of generic, i.e. 'universally' applicable, elements present in any type of dialogue. These occur on different linguistic levels. The first of these levels is the level of syntax, where we are dealing with generic *structural units*, somewhat similar to the sentences we expect to find in written language, but obviously, due to the nature of spoken language, not too similar.

---

[1] http://www.ling.lancs.ac.uk/groups/spaac/SPAAC.htm .

On the next level, we find generic *morphosyntactic (lexical) units*. These represent vocabulary items that tend to recur in any given dialogue, due to their 'universal' nature. It is here that we will probably also find some of the most striking typological differences between English (or any other Indo-European language) and Korean.

At the next level, the level of what could generally be referred to as 'the content level', we can find generic *semantic & semantico-pragmatic units*. In addition to the generic content, we also have domain-specific semantic units.

At the final level, we are dealing with *speech acts*, in the more recent literature sometimes also referred to as *dialogue acts* (Jekat et. al. '95, Alexandersson et. al. '97). These are the generic elements in a dialogue that reflect the intentions or goals of the participants. We will discuss all of these elements in more detail in the following sections.

4. Segmentation Units in Spoken Language

The first questions we need to ask ourselves when segmenting dialogues for analysis are "What are the units we are dealing with?" or "Are traditional sentence types/categories enough?". he latter can be answered by a very emphatic "No!" because the categories used in more traditional approaches to grammar don't cater for short and incomplete utterances, such as

- yes (, please) ||[2] no (, thank you)

or

- right, fine, well, aha, etc.

and are in most cases concerned more with form than with function. We can see that the same features that are true for spoken English are also quite true for Korean if we simply list the equivalents of the above units for Korean:

- 예, 네, 응 || 아니*, 됐어*, 괜찮아*

or

- 그럼, 그러면 , 괜찮아*, 네 그래요, 좋아*.

---

[2] I am here using the double pipe symbol to signal alternatives as is often done in programming languages.

So what should actually be the appropriate unit of analysis for spoken interactions? I propose that the most useful unit would be *C-unit*, a unit which is clause-like and pragmatic, rather than syntactic in its function, i.e. 'unit of meaning'. Biber et. al. (1999: p. 1070) define the C-unit as a grammatical construct comprising both "clausal and non-clausal units […] that […] cannot be syntactically integrated with the elements that precede or follow them." and this is exactly the type of unit we encounter in many dialogues. As I have pointed out above, though, at least some of the potential C-unit types still resemble the traditional sentence categories we are used to from written language, but may just need to be seen from a new perspective.

4.1. The 'C-Unit' Type Taxonomy

The taxonomy of C-units used on the SPAAC project comprises the following units.

4.1.1. Single Yes- and No-like Answers/Statements

English	Korean
yes, please ‖ no, thank you	네 (그렇게해요) ‖ 아니 괜찮아*

These clearly signal acceptance or rejection respectively. Their somewhat more neutral counterparts are

English	Korean
yes ‖ no	네 ‖ 아뇨

, which on the surface signal acknowledgement or negation of a proposition/question, but may also – in certain contexts – signal the same functions as the units given above.

4.1.2. Discourse Markers

Here, we can essentially distinguish two different functions, the first one being similar to the acknowledging function of the yes/no units above, as in

English	Korean
aha, right, fine, ok, etc.	괜찮아*, 그래*

and the second one being an 'initialise' or initiating one, as in

English	Korean
well, now, so	아, 그럼, 있잖아* , 좋아*

4.1.3. Wh-Questions (Question Word Questions)

English	Korean
who, what, when, where, how	누구/누가, 무엇*, 언제, 어디, 어떻게, 얼마, etc.

The main difference here between English and Korean is that question words in Korean seem to be unambiguous, whereas they can be ambiguous in English, e.g. *this is when i'm going to arrive*. In particular in English wh-words often occur as relative pronouns.

4.1.4. Y/N-Questions

English	Korean
do/can you, is there, etc.	-까, -니, *ㄴ데*[3]

As we can clearly see, whereas in English this type of question is relatively easy to identify because of the syntactic inversion of subject and auxiliary, in Korean it is equally easy to spot it because of the occurrence of the question markers. Of course this does not mean that we can identify all yes/no-questions either in Korean or in English in this way. As is well-known, many questions in Korean do not end in question markers, but actually in 요 and thus on the surface appear like declaratives, but this problem is actually one that is not too dissimilar to the one in English, where often questions in dialogues also have a declarative form. I would argue that, in the absence of any prosodic clues that would help to disambiguate these types of units, we will have to try and identify means of disambiguation them via the context for both languages.

4.1.5. Declaratives

As far as proper declarative C-units are concerned, these units are – along with the two question types discussed above – among those that resemble traditional sentences the most. However, in spoken language, I believe that the status of what is commonly termed subordinate structures needs to be re-examined, as I will try to explain after giving some examples below.

English	Korean
you are able to get the next available train	*합한 가격이 칠만 이천 육백원이거든요*

versus

[3] similar to tag-question in English.

English	Korean
if you miss the service i've reserved you on	*안 돼 있으면 (해주세요)*

We could call the first type ‘pure’ declaratives and the second one ‘subordinate’ declaratives, but does this really make sense in light of the fact that if one reverses their order, only the thematic focus seems to shift, rather than one of the two units truly being ‘superordinate’ to the other? It is exactly for this reason that I prefer to treat these as two separate and independent units.

4.1.6. Imperatives

In English, imperatives are relatively easily spottable, provided that one has an appropriate lexicon listing the stems of verb occurring in the dialogue, which is definitely achievable for limited transactional domains. Negative imperatives are even easier to spot because in modern English they’ll always start with *don’t*. There is also another type of imperative which occurs very frequently and signals a suggestion being offered by the speaker – usually starting with *let’s* – or often initiating a ‘holding phase’, as in *let me think.*

For Korean, we partly have more easily usable clues, such as the occurrence of the ending –자 for the equivalent of *let’s* or a single verb stem as the equivalent of the ‘straightforward’ imperative in English. In addition, we also find the ‘polite imperative’ marked by –십시오 or, in more difficult cases, a single verb form ending in –요 (possibly preceded by an adverb).

4.1.7. Fragments

Fragments are all those C-units that remain after ‘elimination’ of all possible other categories. The often include what traditional grammar would consider ‘ill-formed’ sentences, often because they either lack a finite verb form, do not contain a verb at all, because auxiliaries may have been dropped or simply because they are incomplete:

English	Korean
e.g. good afternoon, Sandra speaking, at 10 o’clock, on the half hour, etc.	*더블룸으로요, 칠월 이일부터 팔월 이십오일까지*

5. Establishing C-units for Korean

For English, looking at the first four words – apart from initial fillers or conjunctions – generally allows us to determine the C-unit type quite accurately. For Korean, this approach would not work

because, in most cases, one needs to look at the verb form, i.e. the end of the unit. Therefore, an appropriate methodology is needed which allows to split dialogues into units automatically. This could potentially to a large extent be achieved automatically by breaking turns at:

- appropriate verb endings (-요, -까, -니, -십시오) or verb stem +-고
- conjunctions like 그리고, 그래서, 그리니까, -면, etc.

Of course, the above should by no means be seen as an exhaustive list of options, but only serve to indicate how this particular problem of handling C-units in Korean could be overcome.

6. The Generic Lexicon – Motivation

The development of the generic lexicon during the SPAAC project arose out of a need to analyse different types of more or less domain-specific data, i.e. to be able to switch between different domains. However, when working on materials from different domains, one cannot help but observe that some lexical elements always remain constant, e.g.

- function words, question words
- common verbs , auxiliaries
- terms of address, etc.

As well as being applicable to different domains, this principle may also be seen as applicable to different languages. The main difference between an agglutinating language such as Korean and English as a mildly inflected language is that for Korean many of the identification processes used during syntactic analysis are of a more morphological nature, rather than simply looking up the part of speech of a word, as one could do for English. Some of the items occurring in a lexicon for English, such as for example auxiliaries, would not necessarily need to be included in a lexicon for Korean, but rather written into the morphological rules used for parsing.

Apart from the idea of including the most common words in the generic lexicon, some of the main ideas developed for English may not work as well for Korean, so I will only sketch them briefly.

The first one is that words change meaning in context and that this change in meaning is often a change in function according to domain, e.g. *book* as ‘reading material’ vs. *book* as ‘reserving a seat/ticket’. The second is that this change in meaning is often associated with change in word-class

– something which I refer to as ‘grammatical polysemy’ –, e.g. *book* as N vs. *book* as V. The latter is obviously much less likely and maybe even impossible in Korean, due to its morphological system.

The third and last idea is that some meanings are more generic/prototypical than others, so that e.g. *book* is far more likely to occur as a noun in English, so the generic lexicon should include this information as it may always be changed by adding domain-specific lexical information.

6.1. Determination Strategies

So how can we go about determining generic lexical items? First of all, we can start by isolating ‘pure’ function words (conjunctions, articles, pronouns, quantifiers , prepositions, question words, deictica, fillers, particles). Then, we can isolate other ‘function words’ such as auxiliaries or be-forms and finally determine high-frequency or ‘everyday life’ content words from large-scale corpora, such as the BNC or Sejong Corpus or empirical observation from materials under analysis using intuition/linguistic knowledge.

Once we have identified the relevant vocabulary, we can proceed to determine the most prototypical functions from tagged corpora by comparing tag assignments and using intuition/linguistic knowledge. As will have become clear from my description of the methodology, in order to achieve the right result, one should not proceed either by only using so-called ‘statistical’ information from frequency lists or simply use one’s intuition, but rather opt for a combination of the two. Using only the first method may cause problems because even large-scale corpora may be somewhat skewed and possibly also contain many errors due to low standards of transcription and might possibly not contain the specialised vocabulary needed for the analysis of dialogues, whereas only using one’s intuition is highly error-prone.

6.2. Usage

As the generic lexicon is often not sufficient to analyse the finer details of specific dialogues, it only needs to be set up once, but latter needs to be augmented by one or more domain-specific ones. As many words can have multiple POS categories by default, mark the most prototypical in the generic lexicon , e.g. for English N for pure nouns, n for words that tend to be nouns, etc. or for

Korean 는 as a subject marker vs. 'adjectiviser'[4]. Once the generic lexicon has been compiled, we can go about setting up domain-specific ones, including domain-specific words and domain-specific POS tags, if applicable.

At run-time, we can then combine the lexica for data analysis, adding the domain-specific vocabulary and overriding generic POS tags with domain-specific usage, if necessary. In this way, we can avoid many cases of ambiguity before they may even arise, but of course our analysis routines also need to be written so that they can cope with the 'intentional ambiguity' introduced by using upper and lowercase tags.

7. Identifying Content

When reading the relevant dialogue systems literature, we often find that content is modelled by a process called *topic* or *keyword spotting*. However, this may be a somewhat simplistic view because, often, individual words do not characterise topics enough. It therefore often makes more sense to try and identify 'keyphrases'.

But what actually is the content of a dialogue? And is it purely semantic in nature as some people seem to assume? I would argue that we need to distinguish at least three different levels of content in any given dialogue:

1) semantic content, which I call 'topics'. Here, we can again distinguish between two sub-categories
    a) completely domain specific content, i.e. much less likely to occur across different domains: type of room/ticket, etc.
    b) generic topics, i.e. content that is fairly likely to occur in most types of dialogue: references to times, places, addresses, directions, etc.
2) generic semantico-pragmatic content, i.e. content that reflects 'everyday interaction' or linguistic concepts, i.e. is more pragmatic in nature – 'modes'
3) generic pragmatic content – speech acts

[4] Provided that one does actually choose to include grammatical markers in the lexicon at all.

I will give examples of and explanations for each of the individual levels in the following sections below and explain how they can be used in order to gather all the information needed in order to interpret specific parts of the dialogue.

8. Generic and Domain-Specific Semantic Content (topics)

Topics are semantic items of content that reflect what the dialogue is actually all about, i.e. reflect the subject matter or overall purpose of the dialogue. As pointed out above, we can distinguish between domain-specific topics, that are more or less only likely to occur in dialogues of restricted domains and generic topics, those that are highly likely to recur throughout a variety of different – if not even all – domains. The table below lists examples of domain-specific topics from the two domains of train timetable information and bookings and hotel reservations, where it is perhaps not very surprising that many of the expressions used in Korean are actually taken straight from English.

label	English	Korean
booking	book(?!ed up), debited, reference	예약(?:했.*|해 두었.*| 번호)
departure	depart(ing|ure), leav(e|ing)	떠나.* (train), 체크아웃 (hotel)
fare	advance, cheap, (?<!to)purchase, (?<!to)return, saver, open, pound, single, fare, reduce	미리 예약하.*, 싼 표, 주말 티켓|표, 패키지 상품
room	room, (double|single| standard|twin) (room|bed), suite	룸, 싱글(룸)?, (?:스탠다드|일반)(룸)?, 더블(?:룸|침대), 트윈룸, 스위트
rate	discount, rate, charge, tax, percent	디스카운트, 레이트, (?:룸차지|요금), 룸텍스, 퍼센트
service	service	서비스

Table 1 - domain-specific topics

In contrast to the above examples, the next table provides a list of generic topics, as they may be encountered in many everyday interactions. Many of them represent something akin to ‘cognitive universals’.

label	English	Korean
arrival	arriv(e|al|ing)	도착 (하.*)?
avail	availab(le|ility), booked up	남아 있.*, 없어요
cancel	cancel	취소.*

label	English	Korean
confirm	confirm	확인하.*
creditcard	(credit|debit) card, expiry date, Master card, Visa	신용 카드, 만기일
date	(?<!expiry)date, (?:first|second|third|th) of [ADFJMNOS]	며칠, Sino-Korean number + 일 (.*월)
day	(?:[A-Z][a-z])* (?<!all)day, tomorrow, yesterday y	.*요일, 내일, 어제
address	address, postcode, Avenue	.*시, .*구, .*동, .*아파트
telephone	telephone(?! sale)	전화번호
enum	\d{1,}+, ^\d$	₩d{1,}+, ^₩d$
from (locative)	from [A-Z]	non-temp. expr. + 에서, non-temp. expr. + 부터
location	[A-Z][a-z].+(burgh|by|caster|chester|don|ford|ham|pool|port)	.*산, .*포, .*천
month	January, February, March, ...	Sino-Kor. number+ 월 月
name	initial, name, title	이니셜,이름/성함.*, 스펠링.*, 라스트네임
number	how many, one, two, three, ..., number, once	며칠, Sino-Kor. or Korean Number
spell	(?:alpha|\{letter)	
time	at, time, (?<!good) (afternoon|morning|evening|night), hour, early, late(st)?, o'clock, minute, etc.	시에, 오후, 아침, 저녁, 밤, 시간
to (locative)	(?!<according) to [A-Z]*	non-temp. expression + 까지
week	(last|next) week	(지난|다음) 주

Table 2 - generic topics

As can be seen from the examples, the strings associated with a particular label can be used as regular expressions in order to count how many of these topics occur in a given C-unit and thus to rank its semantic content accordingly. This works in more or less exactly the same way for both

English and Korean, the only difference for Korean being that if part of a syllable is to be matched, the string to be analysed first needs to be 'sequentialised', i.e. split apart algorithmically in order to match individual combinations of jamos.

The applicability of this method actually goes much further than one might anticipate, for example when one compares the regular expressions used for spotting locations, a certain tendency for using landmarks, such as ports, rivers, etc. in the naming conventions of locations can be identified for both languages.

9. Generic Semantico-Pragmatic Content (modes)

Modes are the semantico-pragmatic counterpart to topics. They are even more generic than the generic topics because they reflect the *modus operandi* of particular C-units in a dialogue, i.e. they provide information about specific elements of interaction between the participants and thus are always present in any given dialogue. They represent high-level categories of 'aboutness' and can be categorised into four relatively distinct conceptual fields:

1) grammatical modes
2) interactional modes
3) point-of-view modes
4) social modes

I will give detailed examples and brief descriptions of all four categories below.

9.1. Grammatical Modes

Grammatical modes tend to be represented by conjunctions or expressions of modality, such as possibility, probability, etc. Their importance for particular parts of a dialogue lies in the fact that they signal conditions or circumstances imposed on the participants in the dialogue, which may force these to take particular actions or adopt certain strategies.

label	English	Korean
alternative	either|or	아니면, 이나
condition	if|whether|unless|as long as|while	하면,할 때
constrain	(al)?though|but|only|have (got)? to|must, need	하지만, 해야하.*

exists	there(?:\'s|are), (?:is|are) there	가 있.*
possibility[5]	(?:can|be able|might|may)	할 수 있.*,할 수도 있.* ,혹시
probability	(?:probab|like)ly	아마도 할 것 같*
reason	(?:cos|because|that\'s why)	(?:왜냐하면)? 하니까.*, 때문에.*, 해서.*

Table 3 - grammatical modes

9.2. Interactional Modes

Interactional modes, on the other hand, mainly represent either reactions of one dialogue participant to what the interlocutor has said/asked or 'initiating' *moves*, potentially signalling the beginning of a new phase in the dialogue.

label	English	Korean
back-channel	mh?m	어
intent	i\'ll just, i\'m (?:not)? going to, i\'d like to	할거.*,하고 싶.*
manage	bear with me, hold the line, let me think	잠깐만.*, 기다ㄹ.*
offer	I offer, etc.	에 해줄께.*
preference	prefer, want(?:s|ed)?, wanna, wish, hope, (?:.*d|you) like, (?:i|s?he|they|we) (?:\'d|would) rather, (?:i|we)\'ll go for	좋아.*, (?:그냥)?.*로 하겠어요
reassurance	that\'s (?:ok|fine)	네, 맞아.*
report	i\'m told, i\'ve been told.	라고 들었.*
abandon	…	…
verify	check, consult, look (?:it|this|that) up, verify, i\'ll find out, have a (?:brief|quick) look	확인ㅎ.*, 찾아 보.*

Table 4 - interactional modes

9.3. Point-of-view Modes

Point-of-view modes are constructs that reflect expressions of opinion, ideas or understanding of the dialogue participants. As such, they often incorporate expressions that are traditionally handled

[5] This category is simplified for illustrative purposes. The actual implementation distinguishes between first, second and third person possibility.

under the headings of *knowledge* or *belief*. These concepts, however, assume that it is possible to determine the particular stance and attitude of a dialogue participant with a very high degree of confidence, whereas giving these expressions the status of modes does not represent such a strong commitment.

label	English	Korean
aware-ness	i (?:know|realise|understand), i\'m aware	알아.*, 몰라.*,알겠어.*, 알고있어.*
doubt	i (?:doubt|wonder)(?: if)?	.*르 것같지 않.*, .*르까 해서요
opinion	(?:i|we) (?:think|suppose), belief	(?:내|제) 생각에는

Table 5 - point-of-view modes

9.4. Social Modes

Social modes are relatively self-explanatory. They mainly comprise 'the usual' greetings or goodbyes that are customary for any interaction, as well as 'interpersonal' expressions, such as those of sympathy/empathy, regret or appreciation.

label	English	Korean
apology	apolog(?:ise|y)	사과(?:할께|하겠습니다)
appreci-ate	no problem, that would be (brilliant| correct|fine|great|lovely|wonderful)	좋습니다
thank	thank(?:s|\syou)	(?:감사합|고맙습)니다, 고마와요
greet	(?:hi|hello|good afternoon)	안녕하(?:세요|십니까), 여보세요
intro	Sandra speaking	
bye	good(?:bye)?	안녕히 계세요
closing		알겠습니다
regret	i\'m (?:very)? sorry, we regret	미안(?:해.*|합니다), 죄송(?:해요|합니다)
expletive	oh shit, damn	에이, 아이 참
insult	you (bastard|idiot), (damn|blast) you	

Table 6 - social modes

Although many of the labels given for the modes discussed above may actually look like speech act labels, they should not be mistaken for such, but actually seen as pointers towards the identification of a particular speech act as expressed in a unit. A classic example for this is the occurrence of English *hello* or Korean 여보세요 in the middle of a dialogue, where its function is not that of a proper greeting, but rather the signal of a restarting or uptake of an interrupted dialogue, usually because one of the participants has tried to look up a particular piece of information for the other.

10. Generic Speech-Act Units

10.1. Motivation

The motivation on the SPAAC project for coming up with a taxonomy of a generic speech acts was that most annotation schemes surveyed during an earlier project (see Leech et al. '98) tended to use very domain-specific tagsets which were not easily transferable from one domain to another. The most generic annotation scheme so far was the DAMSL (Dialog Act Markup in Several Layers; Allen & Core, '97), which was still not intuitive enough for our purposes and also still too firmly rooted in some of the older philosophical traditions, probably going back to Searle (1969).

The new SPAAC taxonomy is was devised following some basic, but essential, assumptions:

a) there exists a set of high-level speech-act/interactional categories

b) previous annotation schemes have often conflated too many c-units into single moves

c) the range of possible speech-acts for any C-unit is limited by C-unit type

d) c-unit type + mode & topic attributes can be 'combined into' speech-act

The last item in this list already indicates the methodology applied in determining the speech act for a given C-unit, which will be described below.

10.2. Analysis Steps

The first step in the analysis/annotation process is to determine the C-unit type for each unit. At the same time, though, it is often possible to mark default assumptions, e.g. questions tend to be requests for information, requests for directives, etc. Following that, mode, topic and (surface) polarity information is collected, although the latter is currently not yet used for determining the speech act itself, but just written into the XML tag containing the syntactic and default speech act

information. The next step is find answers to questions/request for directives and numerical *echoes*, i.e. repetitions of digits for credit card numbers, etc., and to the respective speech act attribute. If a declarative or fragment is found immediately following the answer, this is also marked as an *elaboration* to the answer or request for directive. Finally, previously unassigned speech act attributes can be determine or previously assigned ones overridden, based on information obtained by first checking against syntactic information and mode attributes or mode + topic attributes. As a last resort, if there is no mode information present or if the mode attributes provide no clues, the speech act may be determined by using topic attributes only.

Below is a list of all the speech act tags currently used in the SPAAC scheme:

speech act label	**brief explanation**
accept	firmly accepting
ackn	acknowledging/ loosely accepting
answ	answer
answElab	elaboration to answer
appreciate	expressing appreciation; possibly accepting
bye	saying farewell; possibly closing the dialogue
complete	completing a unit begun by another party
confirm	repeating what the other party has said in order to confirm details/common ground
correct	correcting details the other party has given
correctSelf	correcting oneself
direct	giving a directive
directElab	elaboration to a directive
echo	repeating what the other party has said for purposes of verification
exclaim	expressing emotion
expressOpinion	expressing an opinion
expressPossibility	expressing possibility
expressRegret	expressing regret
expressWish	expressing a wish, i.e. potentially a mild form of directive
greet	greeting or potential uptake after a *hold*
hold	asking the other party to wait/hold the line
identifySelf	identifying oneself and/or one's institution

speech act label	brief explanation
inform	conveying general information, or signalling awareness
informIntent	signalling the intention to do something
informIntent-hold	as above, but also asking the other party to ‘hold the line’
init	initiating or initialising a new topic, sub-topic or phase in the dialogue
negate	more neutral counterpart to a refusal
offer	offering
pardon	signalling non-understanding or regret
raiseIssue	identifying an issue/a potential problem
refer	deictic reference, usually giving a time, place, etc. as an answer
refuse	refusing an offer/a proposal
reqDirect	asking for a directive
reqInfo	asking for information
reqModal	a request, which is not clearly classifiable, but contains a modal auxiliary
selfTalk	talking to oneself
suggest	making a suggestion
thank	thanking
thank-bye	thanking + saying goodbye
thirdParty	talking to an external party not directly involved in the current dialogue
unclassifiable	any speech act that does not fit any of the remaining classifications
uninterpretable	classifies a unit that is uninterpretable due to incompleteness or incoherence

Table 7 - the SPAAC speech act taxonomy

11. Conclusion

In this article, I have attempted to show how an automatic annotation scheme originally devised for English can potentially be adapted in order to be applied to Korean dialogues. Although there are probably still many errors and omissions in my presentation due to my still only relatively limited knowledge of Korean grammar, I hope to have demonstrated that this should be possible due to the fact that the methodology employed is a highly cognitive, but at the same time also very surface-oriented one that tries to avoid over-generalisations and inferences that should not be made if the appropriate surface level information is missing.

12. Bibliography:

Allen, J and Core, M. 1997. “Draft of DAMSL: Dialog Act Markup in Several Layers”. available at: ftp://ftp.cs.rochester.edu/pub/packages/dialog-annotation/manual.ps.gz .

Alexandersson, J., Buschbeck-Wolf, B., Fujinami, T., Maier, E., Reithinger, N., Schmitz, B. and Siegel, M. 1997. “Dialogue Acts in VERBMOBIL-2”. VM-Report 204, DFKI GmbH, Stuhlsatzenhausweg 3, 66123 Saarbrücken.

Biber, D., Johansson, S., Leech, G., Conrad, S., Finegan, E. 1999. *Longman Grammar of Spoken and Written English*. London: Longman.

Jekat, S., Klein, A., Maier, E., Maleck, I., Mast, M., Quantz, J. 1995. “Dialogue Acts in VERBMOBIL”. VM-Report 65, DFKI GmbH, Stuhlsatzenhausweg 3, 66123 Saarbrücken.

Leech, G., Weisser, M., Wilson, A., Grice, M. 1998. “Survey and Guidelines for the Representation and Annotation of Dialogue” in: Gibbon, D., Mertins, I., Moore, R. (eds). 2000. *Handbook of Multimodal and Spoken Language Systems*. Dordrecht: Kluwer Academic Publishers.

Searle, J. 1969. *Speech Acts: an Essay in the Philosophy of Language*. Cambridge: CUP.

Weisser, Martin. 2002a. “SPAACy – A Semi-automated Tool for Annotating Dialogue Acts”. *International Journal for Corpus Linguistics*, 8.1.

Weisser, Martin. 2002b. “Determining Generic Elements in Dialogue”. in: *Language, Information and Lexicography* Vol. 12-13. 25th, December, 2003. Institute of Language and Information Studies, Yonsei University. pp. 131-156.

Weisser, Martin. 2004a. “Tagging Dialogues in SPAACy”. Eingereicht an *Traitement Automatique des Langues*.