Focus to Emphasize Tone Structures for Prosodic Analysis in Spoken Language Generation

Lalita Narupiyakul
Faculty of Computer Science, Dalhousie University
6050 University Avenue, Halifax, Nova Scotia, Canada B3H 1W5
Tel. +1-902-494-6441, Fax. +1-902-494-3962
lalita@cs.dal.ca

Abstract

We analyze the concept of focus in speech and the relationship between focus and speech acts for prosodic generation. We determine how the speaker's utterances are influenced by speaker's intention. The relationship between speech acts and focus information is used to define which parts of the sentence serve as the focus parts. We propose the Focus to Emphasize Tones (FET) structure to analyze the focus components. We also design the FET grammar to analyze the intonation patterns and produce tone marks as a result of our analysis. We present a proof-of-the-concept working example to validate our proposal. More comprehensive evaluations are part of our current work.

1 Introduction

A speaker's utterance may convey different meaning to a hearer. Such ambiguities can be resolved by emphasizing accents in different positions. Focus information is needed to select correct positions for accent information. To determine focus information, a speaker's intentions must be revealed. We apply speech act theory to written sentences, our input, to determine a speaker's intention. Subsequently our system will produce a speaker utterance, the result of analysis.

Several research publications, such as (Steedman and Prevost, 1994) and (Klein, 2000), explore prosodic analysis for spoken language generation (SLG). Klein (2000) designs constraints for prosodic structures in the HPSG framework. His approach is based on an isomorphism of syntactic and prosodic trees. This approach is heavily syntax-driven and involves making prosodic trees by manipulation of the syntactic trees. This approach results in increased complexity since the type hierarchy of phrases must cross-classify prosodic phrases under syntactic phrases. Haji-Abdolhosseini (2003) extended Klein's approach. Rather than referring to syntax, Haji-Abdolhosseini sets the information domain to interact between the syntactic-semantic domain and the prosodic domain. His work reduces the complexity of type hierarchies and constraints which are not related to the syntactic structure. He designs the information structure and defines constraints for the HPSG framework. However his work limits the number of tone selections because he only defines two tone marks: rise-fall-rise and fall to annotate a sentence.

Our work is inspired by Haji-Abdolhosseini's work. We design the focus structure for spoken language generation. Based on the focus theory (Von Heusinger, 1999), the focus part identifies what part of the sentence can be marked with the strong accent or emphasized by a high tone. By analyzing speech acts, we can understand how speech with prosody can convey distinct speaker intentions to a hearer. In the next section, we present an overview of our FET (Focus to Emphasize Tone) system and its processes. We will explain how to analyze focus information, design the FET structure, and find the relationships of focus with speech acts to prosodic marks in section 3. We implement our FET grammar for the Linguistic Knowledge Base (LKB) system (Copestake, 2002), generate a set of focus words, explain the FET environment, and show an example in section 4. In the last section, we conclude the current state of our work and the future work.

Proceedings of the COLING/ACL 2006 Student Research Workshop, pages 67–72,
Sydney, July 2006.

2 Overview of FET System for Prosodic Analysis in SLG

Our system generates the prosodic structure depending on the focus analysis. We use this prosodic structure to modify synthetic speech for SLG. Our FET structure is constrained by the speaker's intention. To define prosody, we explore the relationships of focus and speech acts from various sentence types. The diagram of our FET system is shown in figure 1 and we present an overview of the FET system based on the LKB system below.

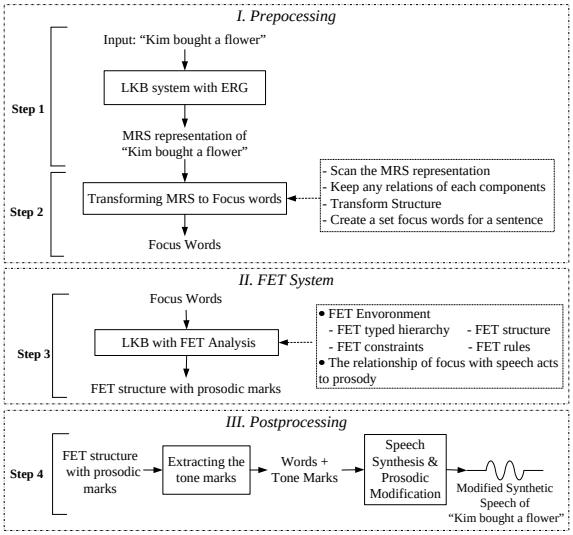

Figure 1: A diagram of the FET system

Our input is a sentence and its focus criterion obtained from a user. In figure 1, the example sentence is "Kim bought a flower" and the focus criterion is G (see table 2). Our system is composed of four main steps.

The first step is preprocessing. The LKB system with the English Resource Grammar (ERG) (Copestake, 2002) parses a sentence. The LKB system analyzes the syntactic and semantic structures and generates the Minimal Recursive Semantic (MRS) (Copestake et al., 1995) representation. This step occurs before invoking the FET system.

In the second step, we scan the MRS structure and collect any components and their relations among them obtained from the preprocessing step. We select only required information, such as sentence mood, from the MRS representation, assign a speech act code referring to a main verb of a sentence, and obtain from the MRS structure a set of focus words. These focus words are an input for the focus information analysis in the FET system.

The third step is the FET analysis. This step generates the prosodic components inside the FET structure. Using our FET grammar, we input the focus words into the LKB system with the FET environment. This environment consists of the FET type hierarchy, constraints, rules, and structures including the focus and prosodic features. Since the LKB system with FET environment can analyze the focus relations corresponding to speech acts and sentence moods, the system completes the FET structure by generating a set of appropriate prosodic structures containing prosodic marks as a result.

The last step is the postprocessing process. We extract words and their prosodic marks as Tone and Break Index (ToBI) representations (Silverman et al., 1992) from the FET structure. The extracting system processes the FET structure, extracts only our required prosodic fields. These fields are a set of words and their tone marks for a sentence. We use the set of words with tone marks to modify synthetic speech, which is generated by speech synthesis. We use the PRAAT (Boersma and Weenink, 2005) to modify the prosody of the synthetic speech for a sentence. Our output is an audio file of the sentence with modified prosody. Modifying prosody follows the tone marks which are analyzed by the FET system.

3 FET Analysis

We describe our concept of the FET analysis (see step 3, figure 1). We determine how the speaker's utterances are influenced by a speaker's intention. Focus information can be used to indicate how to appropriately mark a part of a sentence to convey the speaker's intention. Focus can scope the content in a sentence to which a speaker wants the listener to pay attention. We also consider speech acts which involve a speaker's intention and speaker's utterance. We analyze the relationships of focus parts with speech acts to tone marks. We define the intonation patterns depending on particular focus parts and speech acts. Our FET analysis obtains syntactic and semantic contents from the preprocessing process. We employ the LKB system to parse a sentence. The LKB system is an HPSG parser. A particular grammar, used for LKB system, is called ERG containing more than 10,000 lexemes. The LKB system generates the semantic information which is represented by MRS representation.

3.1 FET Constraints

Our FET analysis uses a constraint-based approach. We find what part (actor, act, actee or their combinations) must be in the focus from the the MRS structure. If the focus is marked at a position in a sentence then the speaker wants the hearer to recognize the content at that position in the sentence. For example, the speaker utters the sentence "Kim bought a flower" by emphasizing at the different positions in the sentence as shown table 1. Then we transform the MRS structures to our FET content structure which is represented by a set of focus words. This structure contains "actor" (a person or a thing that acts something in a sentence), "act" (an activity in that sentence), and "actee" (the response of the activity) parts.

Table 1: The different focuses in the sentence

	Focus	**Speaker wants to focus at ...**
[a]	$[KIM]_F$ bought a flower.	(Who bought a flower?)
[b]	Kim bought [a FLOWER]$_F$.	(What did Kim buy?)
[c]	Kim [BOUGHT a flower]$_F$.	(What did Kim do?)

Considering a focus part, our focus model will acknowledge two focus types: *w-focus*, and *s-focus*. The *w-focus* represents wide focus, which covers a phrase or a word. The *s-focus* represents single focus, which is placed on a word in the sentence. We assign the actor and actee parts as single or wide focus while the act part is only an *s-focus*. Normally, the focus does not cover only the act part. If the focus covers the act part, then the focus must cover at least one of the related parts (actor or actee). Therefore, we set the focus types following all situations that occur and call the focus criteria. Eight focus criteria are shown in table 2.

Table 2: The focus parts and the focus types

No.	Focus Parts	Focus Types
A	actor+act+actee	{w-focus(actor),s-focus(act),w-focus(actee)} or undefined
B	actor+act	{w-focus(actor),s-focus(act)}
C	actor+actee	{w-focus(actor),s-focus(actee)} or {w-focus(actee),s-focus(actor)}
D	actor	w-focus(actor) or s-focus(actor)
E	act+actee	{s-focus(act),w-focus(actee)}
F	act	s-focus(act)
G	actee	w-focus(actee) or s-focus(actee)
H	$\oslash$	undefined

We define constraints to select the focus types following the different situations. We categorize the conditions for focus types to five cases. These conditions cover all possible situations. These situations define the focus based on the focus parts for most simple sentences. We illustrate the attribute value matrix (AVM) structure to represent these situations in figure 2.

(a) *An s-focus of the actor or actee parts.* The last node in the list of objects is defined as the focus position to emphasize tone (*FET-obj*), see figure 2(a).

(b) *A w-focus at the actor or actee parts.* The list of objects is the FET-obj in the sentence as shown in figure 2(b).

(c) *A w-focus at actor or actee parts containing the multiple lists of objects.* The lists are merged together to be the *FET-obj* as shown in figure 2(c).

(d) *An s-focus at actor or actee parts containing the multiple lists of objects.* If the focus type is an *s-focus* and there are *m* sets of lists of objects (multiple lists of objects), then these lists of objects can be split into the *s-focus* of each list of objects, see figure 2(d).

(e) *A focus on the act part.* Two cases of defining the focus types are shown in figure 2(e). The first case, the *s-focus* marks the act part while the *w-focus* marks the actee part. The second case, the *s-focus* marks the act part and the *w-focus* marks at the actor part.

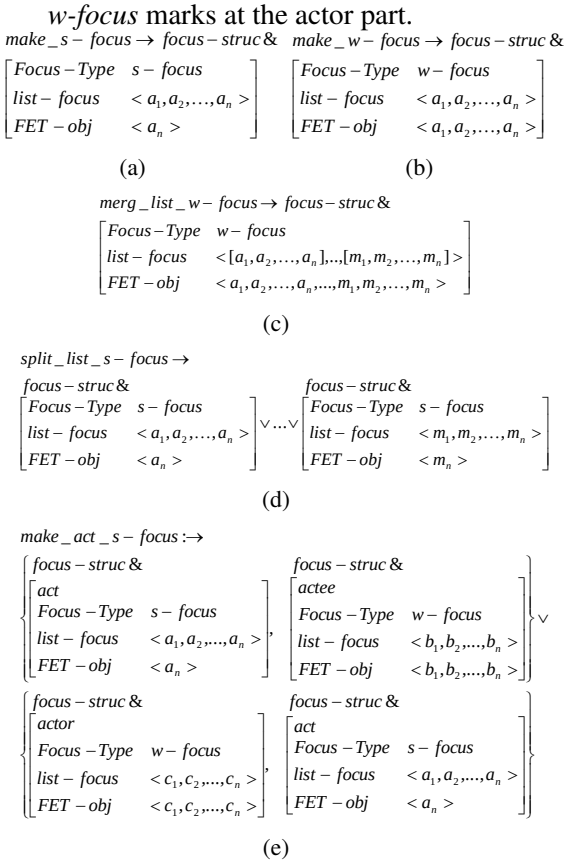

Figure 2: The AVM structure of focus marking: For actor or actee part, (a) *s-focus* (b) *w-focus* (c) *w-focus* of the multiple lists (d) *s-focus* of the multiple lists and, (e) *s-focus* for act part

3.2 The Relationships of Focus with Speech Acts to Prosody

At step 3 of figure 1, we define the speech act codes following Brennenstuhl (1981). To mark

these codes, we consider the main verb (known as the act part inside the FET content structure). These codes define what the speech act categories can be in each sentence. A sentence can be marked by more than one code according to speech act classification (Ballmer and Brennenstuhl, 1981). We mark the speech act codes for 62 sentences from a part of the CMU communicator dataset (2002). Considering the relationships between speech acts and focus parts, we found some common patterns for marking tones in a sentence. For example, the tone mark L-L%, analyzed as low phrase tone (L-) to low boundary tone (L%), is marked at the last word of a sentence for any affirmative sentence. The tone marks H- (high phrase tone) and L- are marked at the last word before conjunction (such as "and", "or", "but", and so on) or are marked at the last word of the current phrase (following the next phrase). We know that the tone mark H* (high accent tone) is used to emphasize a word or a group of words in a sentence. If we want strong emphasis at a word or a group of words then we use the tone mark L+H* (rising accent tone) instead of H*. The groups of speech acts, that we consider in this paper, include intending (EN0ab), want (DE8b), and victory (KA4a), to explore tone patterns. We analyze the relationships of speech acts and tone marks grouping by focus parts as shown in figure 3. Since our example sentence has focus at actee part, speech act code is en0ab, and the sentence mood is affirmative sentence (aff), we define the tone marks for a set of words in the actee part as L+H* L-L%, following figure 3. The outcome of this process is the FET structure including the prosodic structure.

Code	Act Type	Sent Type	Condition
EN0ab	Actee	Aff	$Actee\_tone \leftarrow ([L^* \vee (L+H^*)]+L\text{-})^{n\text{-}1}+([L^* \vee (L+H^*)]+L\text{-}L\%)$
		Int	$Actee\_tone \leftarrow ([H^* \vee (L+H^*)]+H\text{-})^{n\text{-}1}+([H^* \vee (L+H^*)]+H\text{-}H\%)$
DE8b	Actee	Aff	$Actee\_tone \leftarrow H^*+L\text{-}L\%$
		Int	$Actee\_tone \leftarrow H^*+H\text{-}H\%$
KA4a	Actor	Aff	$Actor\_tone \leftarrow H^* \vee (L+H^*)$
		Int	$Actor\_tone \leftarrow H^* \vee (L+H^*)$
	Actee	Aff	$Actee\_tone \leftarrow ([H^* \vee (L+H^*)]+L\text{-})^{n\text{-}1}+L\text{-}L\%$
		Int	$Actee\_tone \leftarrow ([H^* \vee (L+H^*)]+H\text{-})^{n\text{-}1}+H\text{-}H\%$

Figure 3: Tone constraints

4 An Example of FET Implementation with LKB System

In this section, we implement our system using the LKB system with the FET environment. We analyze an example sentence "Kim bought a flower" using the FET system. The system contains the FET environment (see section 4.2) and constrains focus and prosodic features based on FET analysis in section 3. We introduce the FET type hierarchy and describe the components of FET structure.

4.1 Interpreting the MRS representation for Focus Words

In the preprocessing process, the LKB system with ERG parses a sentence and generates the MRS representation (see step 1, figure 1). By scanning each object inside the MRS representation, we keep all reference numbers, mapped with their objects and record every connection that is related to this object and this reference number. We extract only necessary information to generate a set of focus words (see step 2, figure 1). These focus words are generated to correspond to the LKB system. For a sentence, we define a speech act code referring to a main verb and obtain a focus criterion from a user.

Each focus word, as shown in figure 4, is marked by a focus part (*focus-part*). A focus word structure (*focus-word*) contains the focus criterion (*fcgroup*), speech act code (*spcode*), sentence mood (*stmood*) and focus position (*focus-pos*) in a focus part. In figure 4, the focus criterion is defined as group G (see table 2) while the speech acts code is en0ab (intending). The sentence mood referring from MRS is affirmative sentence and focus position is the last node (*ls*). We will describe the *focus-word* and its components in the next section. In figure 4, "Kim" is a actor part while "bought" is an act part. The words "a" and "flower" are the actee parts.

```
Kim := focus-word &
[ ORTH "Kim",
  HEAD actor-part &
       [ AGR1 ls-actor_G-aff-en0ab],
  SPR <  >,
  COMPS < > ].

a := focus-word &
[ ORTH "a",
  HEAD actee-part &
       [AGR1 pv-actee_G-aff-en0ab ],
  SPR < >,
  COMPS < > ].

bought := focus-word &
[ ORTH "bought",
  HEAD act-part &  [ AGR1 ls-act_G-aff-en0ab ],
  SPR < [HEAD actor-part &
         [ AGR1 ls-actor_G-aff-en0ab ] ] >,
  COMPS < focus-phrase & [HEAD actee-part &
         [AGR1 ls-actee_G-aff-en0ab  ]]  > ].

flower := focus-word &
[ ORTH "flower",
  HEAD actee-part &
       [AGR1 ls-actee_G-aff-en0ab  ],
  SPR < [ HEAD actee-part &
       [AGR1 pv-actee_G-aff-en0ab ]] >,
  COMPS < focus-phrase & [HEAD actee-part &
       [AGR1 ls-actee_G-aff-en0ab  ]]> ].
```

Figure 4: A set of focus words

4.2 FET Tone Environment

In FTE system, we provide a set of focus words to the LKB system with the FET environment (see step 3, figure 1). This environment contains the constraints, rules, type hierarchy, a set of features, and their structures for the FET analysis. We design the FET type hierarchy as shown in figure 5. We define three main groups of feature structures: **focus-value**, **prosodic-value** and *feat-struc* to control the focus constraints. **focus-*

*value** represents the focus structures. It is composed of five subfeature structures: focus criterion, focus type (*fctype*), focus name (*focus*), focus position (*focus-pos*), and checking whether a tone mark can be marked at a word (tone-mark). **prosody-value** represents the prosodic structure. Four prosodic subfeature structures are sentence mood, speech act code, accent tone (*accent-tone*), and boundary tone (*bound-tone*). *feat-struc* contains the core FET structure that constrains the relationships between focus and prosodic features. The *feat-struc* structure is composed of six main subfeature structures: (i) focus category structure (*focus-cat*) is a set of constraints which are the combinations of a focus part and a focus criterion such as act_g, actor_g, actee_g, and so on, (ii) focus part structure (*focus-part*) classifies act part and non-act part as actor part or actee part, (iii) focus structure (*focus-struc*) is a subfeature structure of *focus-word* and *focus-phrase*, (iv) checking whether prosodic marks can be marked (*prosody*), (v) prosodic mark (*prosody-mark*) structure maps between types of prosodic mark and accent and boundary tones: no-mark, hEm_Sh-break, etc, (vi) a set of prosodic marks (*prosody-set*) is a set of combinations between accent and boundary tones.

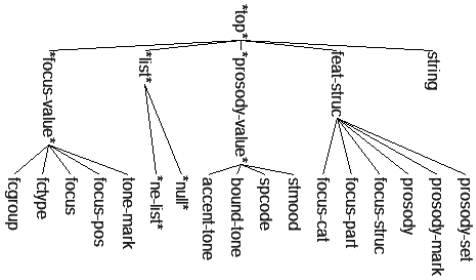

Figure 5: FET type hierarchy

4.2.1 Focus Structure

In figure 6(a), the *focus-phrase* inherits the *focus-struc* with a feature ARGS. The ARGS represents a list of words in the sentence. The focus rules parse the *focus-phrase* with their constraints and define whether tone can be marked at a word in each focus part. The *focus-word* inherits the *focus-struc* with orthography of a word (ORTH) as string. The *focus-word*, as shown in 6(b), represents the focus content structure and corresponds to the LKB system. The *focus-struc*, as show in figure 6(c), consists of HEAD, specifier (SPR) and complement (COMPS) (Ivan et al., 2003). Inside the *focus-struc*, HEAD refers to *focus-part* which is shown in figure 6(d). SPR and COMP are used to specify the components of previous nodes and following nodes in a sentence. Each *focus-part* contains focus and prosodic structures. We classify focus following the possible *focus-cat* for the FET structure. The *focus-cat* controls the constraints for the actor, act and actee parts. The *focus-cat* contains both the focus and prosodic features as a set of subfeatures of the FET structure. This structure contains focus position, focus group, focus type, a set of prosody marks and prosodic structure (*prosody*). The *focus-cat* is shown in figure 6(e).

$$focus-phrase := \begin{array}{l} focus-struc\ \& \\ [ARGS \quad *list*] \end{array} \qquad \text{(a)}$$

$$focus-word := \begin{array}{l} focus-struc\ \& \\ [ORTH \quad string] \end{array} \qquad \text{(b)}$$

$$focus-struc := \begin{array}{l} feat-struc\ \& \\ \begin{bmatrix} HEAD & focus-part \\ SPR & *list* \\ COMPS & *list* \end{bmatrix} \end{array} \qquad \text{(c)}$$

$$focus-part := \begin{array}{l} feat-struc\ \& \\ \begin{bmatrix} FOCUS & focus \\ AGR1 & focus-cat \end{bmatrix} \end{array} \qquad \text{(d)}$$

$$focus-cat := \begin{array}{l} feat-struc\ \& \\ \begin{bmatrix} FOCUS-POS & focus-pos \\ FCGROUP & fcgroup \\ FCTYPE & fctype \\ ADDTONE & addtone \\ PROSODY & prosody \end{bmatrix} \end{array} \qquad \text{(e)}$$

Figure 6: Type feature structure of: (a) *focus-phrase* (b) *focus-word* (c) *focus-struc* (d) *focus-part* (e) *focus-cat*

4.2.2 Prosodic Structure

The prosodic structure consists of these subfeatures: sentence mood, speech act code, and a set of prosodic mark structures. This structure controls the prosodic marks following the FET constraints. These constraints depend on the relationships of focus with speech acts to intonation patterns. The prosody structure is shown in figure 7(a). The accent and boundary tones are mapped with the *prosody-mark* which is illustrated in figure 7(b).

$$prosody := \begin{array}{l} feat-struc\ \& \\ \begin{bmatrix} STMOOD & stmood \\ SPCODE & spcode \\ PROSODY-MARK1 & prosody-mark \\ PROSODY-MARK2 & prosody-mark \end{bmatrix} \end{array} \qquad \text{(a)}$$

$$prosody-mark := \begin{array}{l} feat-struc\ \& \\ \begin{bmatrix} ACCENT-TONE & accent-tone \\ BOUND-TONE & bound-tone \end{bmatrix} \end{array} \qquad \text{(b)}$$

Figure 7: Type feature structure of: (a) Prosodic structure (b) Prosodic mark structure

For focus rules, we have two types of focus rules that are head-complement and head-specifier rules. These rules process the same as a simple grammar rule which is explained in (Ivan et al., 2003). Using these rules, the example sentence "Kim bought a flower" is parsed and the result is the complete FET structure including the focus

and prosodic information. The FET structure of the word "Kim" is shown in figure 8.

```
focus-word
HEAD: [ actor-part
        FOCUS: ACTOR
        AGR1: [ ls-actor\_g-aff-en0ab
              FOCUS-POS: LAST
              FCGROUP: G
              FCTYPE: WS-FOCUS
              PROSODY: [ aff-en0ab-no-prosody
                       SIMOOD: AFF
                       SPCODE: EN0AB
                       PROSODY-SET: [ no-prosody
                            PROSODY-MARK1: [ no-mark
                                 ACCENT-TONE: NOACCENT
                                 BOUND-TONE: NOBOUND ] ]
                            PROSODY-MARK2: [ no-mark
                                 ACCENT-TONE: NOACCENT
                                 BOUND-TONE: NOBOUND ]
                           ] ] ] ] ] ] ] ] ]
SPR: *NULL*
COMPS: *NULL*
ORTH: Kim ]
```

Figure 8: FET structure of the word "Kim"

4.3 Modifying Prosody for Synthetic Speech

In the postprocessing process (see step 4, figure 1), we extract a set of words with tone marks from the FET structure. An example of these words with tone marks is shown in figure 9. Finally we transfer this data to generate the synthesized speech by the speech synthesis and modify prosody.

```
ORTH: Kim                ORTH: a
Focus: actor-part        Focus: actee-part
ACCENT_TONE1: NOACCENT   ACCENT_TONE1: NOACCENT
BOUND_TONE1: NOBOUND     BOUND_TONE1: NOBOUND

ORTH: bought             ORTH: flower
Focus: act-part          Focus: actee-part
ACCENT_TONE1: NOACCENT   ACCENT_TONE1: L+H*
BOUND_TONE1: NOBOUND     BOUND_TONE1: L-L%
```

Figure 9: A set of words and their tone marks

5 Concluding Remarks

We design the FET system based on the small number of sentences from a part of the CMU communicator dataset (2002). These simple sentences relate to traveling information. In this paper, we use the MRS representation from the LKB system to determine actor, act and actee parts. Since the LKB has a limited grammar and produces multiple parses, then we assume that our input sentence can be parsed by the HPSG parser and only a correct output is provided to the LKB system with the FET environment. We analyze the relationships of focus with speech acts to tone marks. To mark tone, we group the tone patterns by speech acts and focus parts. We implement the FET system using LKB and an example is illustrated in section 4 in this paper. Using the LKB with the FET grammar, the system can parse most simple sentences from the CMU communicator dataset and generate the complete FET structure including prosodic marks for each sentence. We are evaluating the FET system with respect to three aspects: appreciation of listeners to tone based on the tone marks from the FET system, conveying the focus content in a sentence to listeners and the correctness of prosodic annotation. In the future, we will finish the evaluations and analyze more relationships of focus with speech acts to tones to support the various sentences.

Acknowledgement

This work is supported by NSERC, Canada, Royal Golden Jubilee Ph.D. program, Thailand Research Fund, Thailand, and King Mongkut's University of Technology Thonburi, Thailand.

References

Mohammad Haji-Adolhosseini. 2003. *A Constraint-Based Approach to Information Structure and Prosody Correspondence*. Proc. of The HPSG-2003 Conf., In Stefan Muller (ed.), CSLI Pub., Michigan State Univ., East Lansing, pp. 143-162.

Ewan Klein. 2000. *Prosodic constituency in HPSG*, Grammatical Interfaces in HPSG. In Ronnie Cann, and Philip Miller, ed., CSLI Pub., pp 169-200.

Copestake A. 2002. *Implementing Typed Feature Structure Grammars*. CSLI Pub., Stanford, CA.

Copestake A., Flickinger D., Malouf R., Riehemann S. and Sag I.A. 1995. *Translation using Minimal Recursion Semantics*. Proc. of the The 6th Int'l Conf. on Theoretical and Methodological Issues in Machine Translation (TMI-95), Belgium.

Silverman K., Beckman M. B., Pirelli J., Ostendorf M., Wightman C., Price P., Pierrehumbert J., and Hirschberg J. 1992. *ToBI: A Standard for Labeling English Prosody*. In Proc. of ICSLP'92, Banff, Canada, pages. 867-870.

Steedman M. and Prevost, S. 1994. *Specifying Intonation from Context for Speech Synthesis*. Speech Comm., 15, 1994, 139-153.

Von Heusinger K. 1999. *Intonation and Information Structure*. The Representation of Focus in Phonology and Semantics. Habilitationsschrift, University Konstanz, pp. 125-155.

Paul Boersma and David Weenink 2005. *Praat: doing phonetics by computer*. Inst. of Phonetic Sciences, Univ. of Amsterdam, Netherlands, http://www.praat.org, Oct. 2005.

Ballmer T. and Brennenstuhl W. 1981. *Speech Act Classification*. A study in the Lexical analysis of English speech activity verbs. Springer Series in Language and Comm., Vol.8. Springer Verlag, New York, 1981.

CMU Communicator KAL limited domain. 2002. Language Technologies Inst., Carnegie Mellon Univ., www.festvox.org, Oct 2005.

Sag, Ivan A., Thomas Wasow, and Emily Bender. 2003. *Syntactic Theory: A formal introduction*. CSLI Pub., Univ. of Chicago Press.