# Labeler Agreement in Transcribing Korean Intonation with K-ToBI

*Sun-Ah Jun[1], Sook-hyang Lee[2], Keeho Kim[3], and Yong-Ju Lee[4]*

[1]Dept. of Ling., UCLA, U.S.A.; [3]Dept. of Eng. Lg. & Lit., Korea Univ.; [2]Division of English & Chinese, Wonkwang Univ.; [4]Dept. of Electrical, Electronics & Information Engineering, Wonkwang Univ., Korea

## ABSTRACT

This paper reports labeler agreement in the transcription of Korean prosody using Korean ToBI (K-ToBI) [9]. Twenty utterances representing five different types of speech were produced by 18 speakers and transcribed by 21 labelers differing in their levels of experience with K-ToBI. Following the stringent metric used for English ToBI evaluation [14,12], consistency was measured in terms of the number of transcriber pairs agreeing on the labeling of each particular word. The results show that for tonal transcriptions of the 32,130 transcriber-pair-words, agreement was 77% for the type of boundaries at the end of each word (i.e., word, AP, or IP), 78% for AP boundaries, and 91% for IP boundaries. For break indices, the agreement score for exact matching in the labeling was 59%, 69% when relaxing the presence/absence of diacritics, and 99% when relaxing within +/-1 level. In sum, the data confirm that the conventions of K-ToBI are adequate, easy to learn, and can be reliably used for research in Korean prosody and for large-scale prosodic annotation in speech databases.

## 1. INTRODUCTION

A framework for the transcription of prosodic information such as tonal events and the degree of juncture between words was developed for English in early 1990s [1]. This framework is known as ToBI (for TOnes and Break Indices), and the reliability of this transcription system and its agreement across labelers has been evaluated [14,12]. Since then, the ToBI transcription framework has been applied to several other languages including German [5], Japanese [15], and Korean [2,9]. However, an evaluation of transcriber agreement and the reliability of these transcriptions have only been done for German ToBI [6] and, on a smaller scale, Japanese ToBI [4,16] (See http://www.ling.ohio-state.edu/~tobi).

In this paper, we report the results of an evaluation of labeler agreement and consistency in the transcription of Korean prosody using Korean ToBI (K-ToBI) [9].

## 2. K-ToBI

The intonational analysis and attendant prosodic model of Seoul Korean adopted for Korean ToBI (K-ToBI) are based on work by Jun [7,8], which was developed by adopting the framework of English intonation proposed by Pierrehumbert and her colleagues [11,3]. A schematic representation of the intonation model of Seoul Korean is illustrated in Figure 1.

Korean has two prosodic units defined by intonation: the Intonation Phrase (IP) and the Accentual Phrase (AP). An AP is smaller than an IP but larger than a word (a lexical item plus a case marker or postpositions), and it is demarcated by phrasal tones. An IP is marked by a boundary tone (X%) and phrase final lengthening. The AP phrasal tones are LHLH when the phrase has more than 3 syllables, but the tones show a rising pattern (e.g., LH, LLH, LHH) when the phrase has fewer than 4 syllables. Additionally, the AP initial tone changes depending on the laryngeal feature of the phrase initial segment. It is H when the AP begins with an aspirated or a tense obstruent (e.g., HHLH), but otherwise it is L.

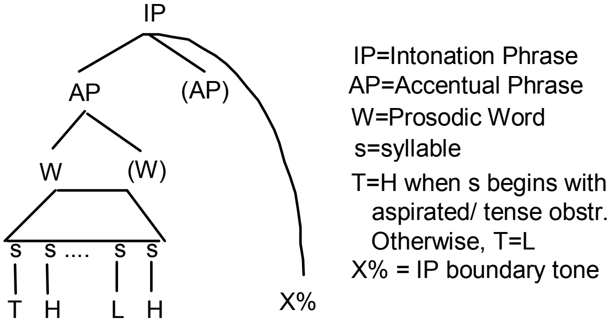

**Figure 1**: Intonation structure of Korean

A first version of K-ToBI was developed in 1994 by Beckman & Jun [2]. It was revised in 1996 to reflect the discussion of a Japanese/Korean working group held during the Prosodic Transcription Workshop at ICPhS in 1995. The evaluation of K-ToBI reported in this paper is based on K-ToBI, Version 3 [9], which was revised from the second version by Jun after a Korean ToBI Workshop in 1998. The third version was presented at the Intonation Workshop at ICPhS in 1999.

According to this most recent version of K-ToBI, a tone tier is sub-divided into two tiers: a phonological tone tier and a phonetic tone tier. This creates five tiers for the transcription of an utterance: a word tier, a phonological tone tier, a phonetic tone tier, a break-index tier, and a miscellaneous tier. **The word tier** corresponds to the 'orthographic tier' in English ToBI. In K-ToBI, a word is considered to be a sequence of segments separated by spaces in written context.

**The phonological tone tier** labels the boundary of two prosodic units: a boundary tone (X%) at the end of an IP and an LHa at the end of an AP. An X% can be one of the 9 IP boundary tones (L%, H%, HL%, LH%, LHL%, HLH%, HLHL%, LHLH%, or LHLHL%). We note that it is possible that not all of the IP boundary tones are distinctive (e.g., LHLH% vs. LHLHL%), but until we find further evidence of distinctive meaning, or lack thereof, we will use all of these tones. **The phonetic tone tier** labels surface realizations of AP tones and the IP boundary tone. The IP boundary tones are the same as those in the phonological tone tier. For AP tones, there are 14 types of surface tonal patterns (LHa, LHHa, LLHa, HLHa, HHa, HLa, LHLa, HHLa, HLLa, LLa, HHLHa, LHLHa, HHLLa, LHLLa). These variations are labeled by three AP initial tones (L, H, +H) and three AP final tones (La, Ha, L+). L and H are the tones of an AP initial syllable, and La and Ha are the tones of an AP final syllable. +H is the tone on the second syllable of an AP, and L+ is the tone on the penultimate syllable of an AP. These six tones are not always realized, but the combination of these six tones can represent all 14 tonal types. For example, a LLH tonal pattern is labeled as 'L' on the first syllable, 'L+' on the penult, and 'Ha' on the final syllable of the AP.

This division of the tone tier was motivated because surface tonal variations are not distinctive or predictable in Korean prosody. Rather, what is distinctive is the phrasing and IP boundary tones. The presence or absence of an AP and IP boundary can change the meaning of an utterance, as with the distinction between wh-questions and yes/no-questions [10] and the disambiguation of syntactically ambiguous structures [13]. Also, the IP boundary tone delivers semantic as well as pragmatic meaning for an utterance.

**The break-index (BI) tier** represents the degree of juncture perceived between each pair of words. There are four break indices: 3 indicates a strong phrasal disjuncture such as an IP, 2 indicates minimal phrasal disjuncture such as an AP, 1 indicates phrase-internal word boundaries, and 0 indicates a juncture smaller than a word boundary. In addition, there are three BI's that indicate a mismatch between the perceived degree of juncture and the tonal pattern. 3m is used when the juncture is 3 but has an AP tonal pattern, 2m is used when the juncture is 2 but has either no AP tonal pattern or the tonal pattern of an IP, and 1m is used when the juncture is 1 but there is an AP tonal pattern.

Finally, **the miscellaneous tier** contains labeler comments concerning events such as silence, audible breathing, laughter, or other disfluencies. The conventions used in K-ToBI for this tier are the same as those of English ToBI.

# 3. THE EXPERIMENT

## 3.1. Speech Data

To assess the labeling conventions of K-ToBI and to demonstrate that these conventions are applicable to various types of speech, we selected twenty utterances representing five different discourse types: TV drama, interview, news, text reading, and story reading. Four sentences were selected from each discourse type. These sentences contained a total of 153 words, and lasted a total of 78.5 seconds. 18 speakers (8 male and 10 female) produced the sentences. Table 1 shows a summary of the speech files:

| Discourse Types | # of Utterances | # of Words | # of Speakers | Total duration (ms) |
|---|---|---|---|---|
| drama | 4 | 31 | 2 male, 2 female | 14,841 |
| interview | 4 | 29 | 2 female | 14,911 |
| news | 4 | 35 | 2 male, 2 female | 16,869 |
| reading | 4 | 28 | 2 male, 2 female | 15,849 |
| story | 4 | 30 | 2 male, 2 female | 16,086 |
| Total | 20 | 153 | 8 male, 10 female | 78,556 |

**Table 1**: Speakers and durations of the five types of speech files

## 3.2. Subjects

Twenty-one labelers, differing in their experience with intonation transcription and in their familiarity with the ToBI model, participated in the experiment. The labelers were divided into four groups: Group 1 (Experts), Group 2 (Familiar with K-ToBI), Group 3 (Familiar with the British intonation model, but new to K-ToBI and intonation transcription), and Group 4 (Beginners, completely new to any model of intonation or prosodic transcription). Each group included five labelers, except for Group 2 which had six labelers. The labelers came from four sites. Six of the labelers from Site A and all of the labelers from Site B and C were provided with 2-3 hours of lecture by the person in charge of each site during which the labeling conventions and background assumptions of K-ToBI were introduced. Site C had 4 hours of group discussion and a review session after the lecture. Two of the labelers from Site A and the one labeler from Site D performed their transcriptions based on the K-ToBI manual alone. Table 2 shows the distribution of labeler groups at each site.

| Sites | # Experts (G1) | # Familiar to K-ToBI (G2) | # Familiar to other model (G3) | # Beginners (G4) |
|---|---|---|---|---|
| A | 3 | 3 | 1 | 1 |
| B | 1 | 1 | 3 | 2 |
| C | 1 | 2 | 0 | 2 |
| D | 0 | 0 | 1 | 0 |
| Total | 5 | 6 | 5 | 5 |

**Table 2**: Distribution of labelers at each training site

## 3.3. Procedure

We selected 20 speech files from the data bank, Phonetically Ballanced Sentences, developed by Wonkwang Univ. in Korea. The first author sent the following materials to the person in charge at each site: 1) the speech files in wave format, 2) the K-ToBI manual, version 3, together with the example sentence files mentioned in the manual, and 3) a Hangul file in which the sentences from each speech file were written in Hangul orthography with empty spaces below for writing tones and break indices. The wave files and the Hangul files for writing transcriptions were necessary because not all sites used the same speech analysis software (they ranged from *xwaves* to *PitchWorks*, *CSL* and *Multispeech*). Each labeler was provided with a copy of the K-ToBI manual and the Hangul file and was asked to use their own software to transcribe two tiers—the phonetic tone tier and the break index tier. We did not ask labelers to transcribe a phonological tone tier because the information in this tier, i.e. AP and IP boundary, can be extracted from the phonetic tone tier. Labelers were encouraged to discuss examples in the manual with others, but not the transcription sentences. After they completed the transcription, their Hangul files were collected and statistics for labeler agreement were applied to the data.

Following the stringent metric for English ToBI evaluation [14, 12], inter-transcriber consistency was measured in terms of the number of transcriber pairs agreeing on the labeling of each particular word. As described in [12], "transcriber pair-word agreement is a stringent metric because when three of four transcribers agree on a label, agreement of that label is reported to be just 50% because only three of the six pairs drawn from the set of four transcribers agree". There are a total of 32,130 pairs for comparison in our data—210 comparison pairs for each word (from 21 labelers) and a total of 153 words.

# 4. RESULTS

## 4.1. Agreement for Tones

Each word may end with an AP final tone (i.e., Ha or La), an IP boundary tone (X%), or nothing at all (i.e., when the word is AP medial). This means that the tones for each word may vary: a word may have no tone, part of an AP tone (e.g., L; H; L +H; L +H L+), all of the AP tones (e.g., L Ha; L +H La; L L+ Ha; L +H L+ Ha), or AP tones plus an IP boundary tone (e.g., L+ H%; H +H L%; L +H L+ LHL%).

The results for tonal transcriptions are shown in Table 3 and 4. Table 3 shows results for Group 1 and for all of the labelers (Group 1+2+3+4). Table 4 shows results from Group 1+2 and from Group 1+2+3. Of the 32,130 transcriber-pair-words, agreement was 77.3% for all labelers for the type of boundary at the end of each word (i.e., AP-medial word, AP-final word, or IP-final word). Labeler agreement for AP boundaries was 77.5% and agreement for IP boundaries was 90.9%. For experts (Group 1), agreement was 81.6%, 81.9%, and 90.9%, respectively. Agreement for the surface realization of AP tones was also examined. For type of word-initial tone, agreement was 82.3% for all labelers and 90.7% for experts. Agreement on the type of word final tone was 74.9% for all labelers and 82.0% for experts. Agreement for the whole tonal pattern for each word, however, was only about 36% for all labelers and 52% for experts. This low agreement seems to be due to the nature of the tonal pattern. That is, there are 14 possible AP tonal patterns and these surface variations are not meaningful or phonological. Furthermore, there is a gross similarity among some of the tonal patterns. For example, rising tonal patterns such as LH, LLH and LHH are very similar to one another, and falling patterns such as LHLL and LHL are also very similar to one another.

| Category | Group 1 | Group 1+2+3+4 |
|---|---|---|
| whole pattern | 52.2% | 35.8% |
| final tone | 82.0% | 74.9% |
| initial tone | 90.7% | 82.3% |
| boundary type | 81.6% | 77.3% |
| AP boundary | 81.9% | 77.5% |
| IP boundary | 90.9% | 90.9% |

**Table 3**: Agreement for tones by Group 1 (Experts) and Group 1+2+3+4 (all 21 labelers).

| Category | Group 1+2 | Group1+2+3 |
|---|---|---|
| whole pattern | 45.0% | 37.5% |
| final tone | 79.4% | 74.8% |
| initial tone | 86.1% | 81.6% |
| Boundary type | 81.2% | 77.8% |
| AP boundary | 81.3% | 78.0% |
| IP boundary | 92.1% | 91.0% |

**Table 4**: Agreement for tones by Group 1+2 and by Group 1+2+3.

Agreement for IP boundaries amongst Group 1+2 labelers is higher than that for Group 1 labelers, but agreement on the detailed surface tonal pattern is higher for Group 1 than for Group 1+2. The agreement results for Group 1+2+3 are very similar to those for Group 1+2+3+4. In sum, the data confirm that the tonal conventions of K-ToBI are adequate for different speech types, easy to learn, and can be reliably used by researchers with different backgrounds.

## 4.2. Agreement for Break Indices

In K-ToBI, each word ends with a break index: 0, 1, 2, 3, #m, #p, or #-. Since BI 3 is predictable at the end of an utterance, agreement for BI transcription was calculated excluding utterance final BI 3. Table 5 shows the results for Group 1 (Experts) as compared to all of the labelers combined, and Table 6 shows the results from Group 1+2 and Group 1+2+3. For the 32,130 transcriber-pair-words, an agreement score of 58.5% was obtained for all of the labelers combined when exact agreement in the labeling of all BI's was examined. The agreement score rose to 68.8% when the presence/absence of diacritics (m, p, -) was not considered, and, finally, agreement reached 98.5% after relaxing the definition of agreement to include transcriptions that were within +/-1 level from one another. Agreement for break indices among experts was 65.5%, 77.1%, and 99.0%, respectively. The results are, in general, close to those for English ToBI [12] (66.6%, 70.4%, 92.5%, respectively) with somewhat lower results for exact matching. Relaxing to the +/-1 level criterion results in higher agreement for Korean than for English most likely because Korean has 4 levels (0-3) of BI's, while English has 5 (0-4).

Agreement for BI's amongst Group 1+2 and amongst Group 1+2+3 falls between the range for Group 1 and that for all labelers. This was expected based on the labelers' familiarity with K-ToBI, though the same trend was not found for tonal transcription. In sum, the data confirm that the break index conventions of K-ToBI are also adequate for different speech types and can be reliably used by researchers with different background.

| Category | Group 1 | All labelers |
|---|---|---|
| Exact match | 65.5% | 58.5% |
| Relaxing diacritics | 77.1% | 68.8% |
| Within +/- 1 | 99.0% | 98.5% |

**Table 5**: Agreement for Break Index for Group 1 and for all labelers combined.

| Category | Group 1+2 | Group 1+2+3 |
|---|---|---|
| Exact match | 63.4% | 58.9% |
| Relaxing diacritics | 74.7% | 69.3% |
| Within +/- 1 | 99.0% | 98.5% |

**Table 6**: Agreement for Break Index for Group 1+2 and for Group 1+2+3.

## 4.3. Agreement/Consistency across Labelers

To ascertain whether disagreement amongst labelers was evenly distributed, we measured percent agreement on tonal elements and break index for each pair of labelers, following [12]. That is, each labeler's transcriptions were compared with each of the other 20 labelers' transcriptions, generating the agreement measures shown in Table 7. Since what is meaningful in Korean prosody is phrasing, we compared word final boundary types and their tonal types. The comparison of tonal elements is divided into three categories. Tone I in Table 7 shows agreement for word final tones plus a boundary type (i.e., AP or IP or None). Tone II shows agreement for word final boundary types only, and Tone III shows agreement for word final tone types only (i.e., L, H, or None). Agreement for break indices within +/-1 level is shown under the BI column in Table 7.

| Labeler ID | Tone I | Tone II | Tone III | BI |
|---|---|---|---|---|
| 1 | 70.4% | 80.9% | 83.7% | 99.1% |
| 2 | 64.4% | 76.2% | 81.6% | 97.9% |
| 3 | 67.3% | 80.7% | 82.0% | 98.8% |
| 4 | 71.7% | 81.7% | 85.4% | 99.2% |
| 5 | 61.1% | 75.8% | 79.1% | 98.9% |
| 6 | 61.4% | 76.9% | 82.5% | 99.2% |
| 7 | 67.9% | 80.8% | 84.0% | 98.0% |
| 8 | 67.7% | 79.9% | 82.0% | 98.9% |
| 9 | 67.7% | 81.1% | 83.2% | 98.8% |
| 10 | 64.1% | 76.5% | 82.8% | 98.9% |
| 11 | 60.7% | 77.8% | 75.8% | 98.9% |
| 12 | 70.1% | 80.6% | 83.5% | 98.4% |
| 13 | 57.6% | 73.8% | 76.3% | 97.1% |
| 14 | 61.5% | 74.7% | 79.4% | 98.9% |
| 15 | 50.8% | 70.8% | 74.6% | 98.5% |
| 16 | 50.4% | 71.8% | 67.8% | 98.1% |
| 17 | 63.7% | 78.9% | 82.3% | 98.5% |
| 18 | 53.5% | 73.6% | 75.3% | 98.3% |
| 19 | 62.3% | 75.9% | 81.4% | 99.4% |
| 20 | 59.3% | 72.3% | 81.4% | 96.8% |
| 21 | 69.8% | 81.5% | 84.0% | 98.9% |

**Table 7**: Percent agreement for each labeler. Labeler ID= labeler's identification number; Tone I= word final tone plus boundary type; Tone II= word final boundary type; Tone III= word final tone type, and BI= break index within +/- 1 level.

Results show a similar degree of agreement across labelers, except for labelers 15 and 16. These two labelers are from Group3, meaning that they are familiar with the British model of intonation. It was found, though, that these labelers actually had little experience working with intonation based on digitized speech data. Labelers 1-5 belong to the Expert group, but their results are not much higher than results for the other labelers. In sum, this suggests that transcribing intonation using K-ToBI is easy to learn if one has some experience with digitized speech data. It also suggests that the training materials are sufficient for learning the conventions even as a beginner.

## 5. CONCLUSION

In sum, the data confirm that the tonal conventions of K-ToBI are adequate, easy to learn, and can be reliably used for different speech types. The data also suggest that K-ToBI could be used as a prosodic annotation system for large-scale speech databases and for research in Korean prosody.

A further study is needed to determine which surface tonal patterns for AP's are confused with which other patterns, and to determine how to incorporate the linking between tone labels and break indices. Finally, a large-scale database annotated using K-ToBI is needed in order to investigate which tonal patterns are the most common and distinctive so that we can resolve our tonal inventory for K-ToBI.

## 6. REFERENCES

1. Beckman, M. & Elam, G. A. "Guidelines for ToBI transcription", version 2.0. Ohio State Univ., 1994.
2. Beckman, M. & Jun, S.-A. "Guidelines for K-ToBI labeling", version 2, ms. OSU & UCLA. 1995.
3. Beckman, M. & Pierrehumbert, J. "Intonational structure in Japanese and English", *Phonology Yearbook* 3. 255-309. 1986.
4. Campbell, N. "Autolabeling Japanese ToBI. In *Proc. ICSLP*, 2399-2402, Philadelphia. 1996.
5. Grice, M. & Benzmuller, R. "Transcription of German using ToBI Tones - The Saarbruken System. *Phonus* 1. 1995
6. Grice, M., Reyelt, M., Benzmuller, R., Mayer, J. & Batliner, A. "Consistency in transcription and labeling of German Intonation with GToBI". In *Proc. ICSLP*: 1716-1719. 1996
7. Jun, S.-A. The *Phonetics and Phonology of Korean Prosody*. Ph.D. dissertation. Ohio State Univ., 1993. [Published by Garland, New York, 1996]
8. Jun, S.-A. "The Accentual Phrase in the Korean prosodic hierarchy", *Phonology*. 15.2. 1998.
9. Jun, S.-A. "K-ToBI (Korean ToBI) labeling conventions: version 3", *The Korean Journal of Speech Science*, Vol. 7, No.1:143-169, 2000.
10. Jun, S.-A. & Oh, M. "A Prosodic analysis of three types of wh-phrases in Korean". *Language and Speech* 39. 37-61. 1996.
11. Pierrehumbert, J. *The Phonology and Phonetics of English Intonation*. MIT PhD Dissertation, 1980.
12. Pitrelli, J., Beckman, M., & Hirschberg, J. "Evaluation of prosodic transcription labeling reliability in the ToBI framework". In *Proc. ICSLP*, 123-126, Yokohama, 1994.
13. Schafer, A. & Jun, S.-A. "Effects of Accentual Phrasing on Adjective Interpretation in Korean*", East Asian Lg. Processing*. CSLI. To appear.
14. Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M. Wightman, C., Price, P. Pierrehumbert, J. & Hirschberg, J. "ToBI:A standard for labeling English prosody," *Proc. ICSLP*. 867-870. 1992.
15. Venditti, J. "Japanese ToBI Labeling Guidelines. [http:/ling.ohio-state.edu/Phonetics/J_ToBI/jtobi_homepage.html], 1995.
16. Venditti, J. "The J_ToBI model of Japanese intonation", in Jun, S.-A. ed. *Prosodic Typology and Transcription* (temporary title). In preparation.

## ACKNOWLEDGMENTS

We are very grateful for the help of the following contributors who participated as labelers: Taehong Cho, Ji-Hyun Jun, Hyun-Sook Kang, Hee-Sun Kim, Hyo-Sook Kim, Min-Jung Kim, Young-Hee Kim, Sahyang Kim, Sun-Cheol Kim, Eun-Jung Kong, Hoyong Lee, Hyuck-Joon Lee, Mira Oh, Mee-Jeong Park, Hee-Kwan Rhyu, Cheol-Jae Seong, Dong-Hyun Shin, Yoon-Kyung Song, Soo-Yeon Yoon. We are also grateful to Illam Park for his help in writing programs to evaluate labeler agreement, and to Jong-Jin Kim for his help in data selection. The speech data used for this experiment and the K-ToBI manual can be obtained by contacting the first author (jun@humnet.ucla.edu).