Social and Linguistic Speech Prosody

Proceedings of the 7th international conference on Speech Prosody

SpeechProsody 7

(Trinity College Dublin) May 20-23, 2014

ISSN: 2333-2042

Intonation Unit Size in Spontaneous Hebrew: Gender and Channel Differences

Vered Silber-Varod [1], *Tal Levy*[2]

[1] The Research Center for Innovation in Learning Technologies, The Open University, Israel
[2] Department of Electrical and Computer Engineering, Rutgers University, USA
vereds@openu.ac.il, tal.tl.levy@rutgers.edu

Abstract

In this corpus-driven research, the question of whether there is a tempo at the Intonation Unit (IU) level, and whether defined IUs differ not only with regard to their pitch contour and boundary tones but also with respect to their phonological size. For this reason, the inventory of syllable size (in terms of segments (phonemes)) and word size (in terms of syllables) was examined, and then each IU category (mainly Terminal vs. Continuous) was measured with respect to the number of syllables and words it contains. Moreover, terminal IU size was also measured with regard to the amount of embedded continuous IUs. Results showed that terminal IUs in spontaneous Israeli Hebrew (IH) do not necessarily consist of embedded continuous IUs. This can be explained due to their massive use as short feedback units in spontaneous speech. Statistical measurements for gender and channel (Face-to-Face vs. telephone conversations) variables were carried with no significance for gender, but with statistically significance for several channel aspects. Last, estimated durational measurements of the IU size are presented.

Index Terms: prosodic unit size, Israeli Hebrew, gender, channel, duration

1. Introduction

When studying speech rate, or speech rhythm, the overall goal is to characterize speech in terms of how many speech units (syllables, prosodic words, etc.) are uttered within a certain time unit (milliseconds, seconds, minutes, etc.). This measure takes into account silent pauses and other types of disfluencies. For example, Syllables-per-second (SPS) is a well know measure for this purpose. For the same reason, analyzing speech unit size in terms of amount of syllables, words, and duration per a certain prosodic unit higher in the hierarchy, can reflects language, or speaker, characteristics. Indeed, since the duration of syllables varies according to their inner-structure [1, p. 55], speech rate is known to be language-dependent, or at least language-type dependent (syllable-timed vs. stress-timed; syllable-rhythm vs. word-rhythm [1, pp. 54-57]. Moreover, in order to characterize prosodic unit at all levels, from syllable to utterance, the prosodic representation "should take into account the heterogeneity and the variations in prosodic constructions encountered in ordinary speech" [2]. There are many applicative aspects of speech rate, such as diagnosing speech disorders, or developing natural speech synthesis. Thus "Obtaining normative data on speaking rate for various groups of speakers is required." [3, p. 131]. For Hebrew, [3] reported that "presently there are no published studies that have directly examined speaking rate among adult Hebrew speakers." [3, p. 131]. Indeed, [3] is a preliminary attempt to provide such data, by quantifying speaking rate within a specific subgroup of speakers – radio newscasters. Nevertheless, it is suggested that speech rate is affected by the communicational setup, specifically, the number of conversation partners was shown to influence rate, such that talking to a single interlocutor is typically performed at a relatively slower rate [4], [5]. Therefore, an examination of the channel aspect as a comparison between Face-to-Face dialogues and Telephone conversations serves this research end. As for gender differences, although there is no linguistic reasoning for such a difference, and indeed no gender differences were found before in Hebrew speech [6], albeit [6] examined the articulation rate and not the speech rate (the latter includes durations of pauses, disfluencies etc.), it was decided to look at this aspect as a baseline examination (and also to use it as a counter-evaluator for the prosodic annotation and segmentation). In the method section (section 2), the prosodic units under investigation are described. In section 3 the analyzed data is described, not only in terms of the corpus type, but also in terms of word and syllable structures. In section 4 the results of gender and channel comparisons are given, and section 5 examines the results and unfolds future research plans.

2. Method

2.1. The prosodic units under investigation

The present paper leans on the prosodic data documented in [7], where the main prosodic unit under investigation is the intonation unit ((IU). For IU definition, see [8, pp. 8-15]. The prosodic annotation process imposes two main types of IU boundary tones that can be defined according to the communicative value of intonation: Terminal (T-) boundary tones and Continuous (C-) boundary tones. A boundary tone was annotated as *T* when the surface intonation signaled that the speaker had "nothing more to say". A boundary tone was annotated as *C* whenever the final tone of the intonation unit signaled "more to come". The recordings were perceptually annotated with a set of prosodic boundaries. Additional validity was achieved using acoustic measurements (see details in [7, pp. 46-55]. This resulted in a back and forth annotation method between perceptual annotation and acoustic measurements, with priority given to major perceptual differences. However, it should be noted that prosodic boundaries were annotated independently of the syntactic structure. The label inventory of prosodic boundaries is as follows. Within the above two types (T- and C- boundary tones), there were two T-boundaries: Terminal Finality (T.), and Terminal Question or Appeal (T?), and five C-boundaries, which were determined according to their tone at the last syllable of the IU: Neutral (C→), Elongated (C:), Rise (C↑), Rise-Fall (C↑↓) and Fall (C↓). Fragmented, or truncated (TR), boundaries were also used in the annotation process.

2.2. The parameters under investigation

In the current research, the following data was extracted for each speaker:

1. No. of words per speaker
2. No. of syllables per speaker
3. No. of pauses (#) per speaker

4. No. of IUs per speaker
5. IU size (word): no. of words per IU
6. IU size (syllable): no. of syllables per IU
7. T- unit size: no. of C-units per T-unit

This will be demonstrated on (1):

heXlaft et ze **T? #** lo ani **C→** lo jodaat ma jihje **T.** (1)

changed-PAST.2SG.F Acc. this **T? #** no I **C→** no know.PARTICIPLE.F what be-FUT.3SG **T.**

'you have changed it**?** no I**,** don't know what is going to be**.**'

For the chunk presented in (1) the following data was extracted:

1. No. of words: 9
2. No. of syllables: 14
3. No. of pauses: 1
4. No. of IUs: 3 (T?, C→, T.)
5. IU size (word): 3/T?, 2/C→, 6/T.
6. IU size (syllable): 4/T?, 3/C→, 10/T.
7. T-unit size (by C-units): 0/T?, 1/T.

The first four parameters are general ones and reflect the speech activity of each speaker.

1. The word parameter counts how many words were transcribed per speaker. This included clitics (function words) that were transcribed separately of their host (the following NP). For example, [ha yeled] 'the boy' was calculated as *two* words and not as a single word, הילד, which is the way the Hebrew orthographic system does.
2. The syllable parameter counts the number of vowels per speaker, following the obligatory presence of a nucleus in the syllable and fact that in Hebrew the nucleus is only a vowel (vowel hiatus, such as in [Raa] see.PST3M 'he saw', was considered as two syllables, but not in cases of word-final diphthongs, where a vowel occurs before the sequence [aX], as in [RuaX] 'wind' (32 such diphthong cases in the corpus)).
3. The pause (marked #) parameter counts pauses above 100ms, per speaker.
4. The IU parameter counts how many IUs (Ts and Cs) were annotated per speaker.
5. IU size (word) parameter counts words per IU, as follows: T-unit size was defined as the number of words from one T-boundary to the next T-boundary. On the other hand, C-unit size was defined as number of words from *any* prosodic boundary to the given C-boundary.
6. IU size (syllable) parameter counts syllables per IU, as follows: T-unit size was defined as the number of syllables from one T-boundary to the next T-boundary. On the other hand, C-unit size was defined as number of syllables from *any* prosodic boundary tone to the given C-boundary.
7. T-unit size parameter refers to the number of minor C-units in major T-units. This parameter reflects rhythm at the intonational unit level.

3. Data

The corpus of this research consists of 19 spontaneous Israeli Hebrew dialogues extracted from The Corpus of Spoken Israeli Hebrew [9]. The recordings were made during 2001-2002. The total duration of analyzed speech is almost five hours. Total number of word types: 4,374; Total number of word tokens: 32,175. The 19 recordings consist of private spoken dialogues in two channels: direct, face-to-face (F2F), dialogues, and distance, telephone (TEL) conversations. Each recording consists of conversations between one core speaker (the "informant", who had the recording equipment on his body for 24 hours) and various interlocutors with whom the speaker interacted on that day. The TEL sub-corpus contains spontaneous phone conversations recorded by the same19 informants. These conversations were part of the 24 hours routine during which the participants recorded themselves. It should be noted that the TEL sub-corpus consists only of the informants' speech, and not their interlocutors'. The total duration of the F2F sub-corpus is 206 minutes; the total duration of the TEL sub-corpus is 83 minutes. Within these 19 recordings, a total of 62 speakers (28 men and 34 women) were transcribed and annotated. The women speech consists of 19,903 words, while the men speech encompasses 12, 272 words. The corpus is heterogeneous in terms of amount of speech per speaker, ranging from 3 to 2,074 words per speaker in the Women sub-corpus; and from 2 to 1,531 words in the Men sub-corpus. The channel groups are also varied. There are only 9 TEL conversations (4 men; 5 women; total of 7,230 words), with speech material ranging from 233 to 2,074 words per speaker; and 61 F2F dialogues (33 women, 28 men; total of 25,107 words), with speech material ranging from 2 to 1,531 words per speaker.

4. Results

The aim of this preliminary research was to investigate if there are: 1. Gender differences; or 2. Channel differences regarding IU size. The data was statistically measured by Mann-Whitney Test, which is a method which has more efficiency on data with non-normal distribution.

4.1. Syllable and word size

In order to measure IU size in terms of words and syllables, it is important to know what the *word* unit size is (in terms of syllables), and what the *syllable* unit size is (in terms of segmental (phoneme) content), in the corpus.

4.1.1. Word size

The statistics of the *word* lengths (in terms of the number of syllables) for the whole database is as follows: The minimum word length is 0 syllables (25 word types; 0.34% in the corpus). This includes cases of interjection such as [m:] 'mm' or truncated words (i.e., false starts). The maximum word length is of 6 syllables. This occurred only once in the loan word [otobijogRafja] 'autobiography'. The average syllables per word (SPW) is 2.3 (standard deviation 0.8). Figure 1 summarizes these statistics, which are close to other studies on word structure in Hebrew child directed speech [10]. Note that in Figure 1 it is demonstrated that monosyllabic words are mostly used in spontaneous IH speech (47.96% tokens), but disyllabic words are four times more varied than monosyllabic ones (black histograms reflects word types). This is an indication to the transcription method carried in the current research, where monosyllabic clitics (i.e., function words such as preposition, definite article, subordinate article) were transcribed as a single orthographic unit, and thus are the most frequent in quantity, but less varied in terms of word-types. Word length statistic was also carried with regard to phoneme per word (PPW). This can be an indication to types (and their relative frequency) of syllable structures in IH. The minimum word-length in terms of number of segments is 1 (18 word-types; 2.47% of corpus). These words include lexemes, such as [o] 'or', interjections and false starts which consist of consonants only. The maximum word length is 13, which

occurred only once in the loan word [otobijogRafja] 'autobiography'. The average word length is 5.6 PPW (standard deviation 1.7).

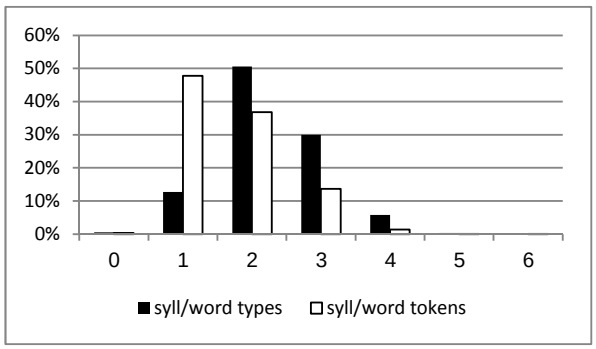

Figure 1: *Syllables per word (SPW) ratios (Types vs. Tokens).*

4.1.2. Syllable size

The average *syllable* length was also calculated using the following method: number of segments per word divided by the number of vowels (i.e., syllables) of that word. For example, in the word [otobijogRafja] 'autobiography' there are 6 syllables. According to the syllable length calculations the syllable length is 2.166 (phonemes per syllable). This is very close to the real average calculation, which is 2, as can be shown from the syllabic division [o.to.bi.jo.gRaf.ja]. There are 4 syllables with 2 phonemes; 1 syllable with 3 phonemes and 1 syllable with a single vowel. By excluding the 25 cases of consonant-based false-starts, the minimal syllable length is 1, meaning the syllable consists of a single vowel, and is related to five different vowel-only words (including [a] and [e] which are interjections). These vowel-only words constitute 2.2% of the corpus. The maximal syllable length (in terms of segments) is 6, and is referred to a single case of the word [dZoRdZ] 'George', which reflects the transcription method of 2 consonants to symbolize the affricate [dʒ]. The average syllable length is 2.6 (phonemes per syllable. standard deviation is 0.7). To sum up, IH speakers speak on average 2.3 syllables per word, an average of 5.6 phonemes per word.

4.2. IU size: gender and channel comparison

In this section, comparison between the two categories: gender and channel, will be analyzed as follows: a comparison with regard to the distribution of IU types; a comparison with regard to IU size in terms of word; Finally, an estimated durational measurements of IUs will be presented.

4.2.1. Gender

In the gender category, no statistically significant differences with regard to IU size were found. The distribution of prosodic boundary tones among females and males is shown in Figure 2. Both groups use the (T.) tone widely, and (T?) to a lesser extent. The T-units' sizes vary from 1-58 words (0-94 syllables). No significant gender differences were found with regard to number of words or syllables, but statistically significant results were found with regard to the gender use of T-boundaries (T and TQ) and C-boundaries ($x^2 = 31.8631; p < .0001$). Figure 2 demonstrates that men use more T(.) and T(?) units than women, who use relatively more C-units (32% of all units, compared to 26% in men's speech). This can imply more intonational variations in women's speech.

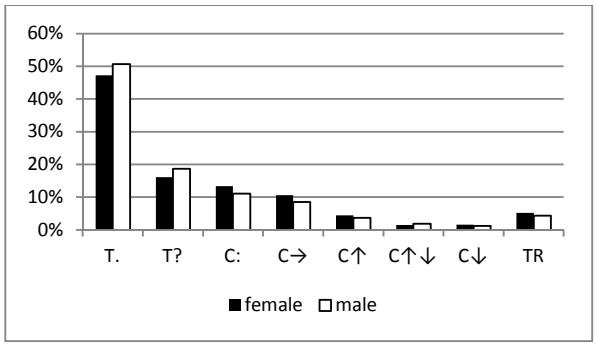

Figure 2: *Distribution of prosodic boundary tones among females and males.*

4.2.2. Channel

The distribution of the prosodic boundary tones in F2F and TEL are shown in Figure 3. The two variables, channel and boundary type (T- or C-) were tested in Chi-square statistics and the results are statistically significant ($\chi^2 = 98.116; p < .0001$), meaning that the two variables show contingency. Moreover, C:-units are relatively more common in TEL.

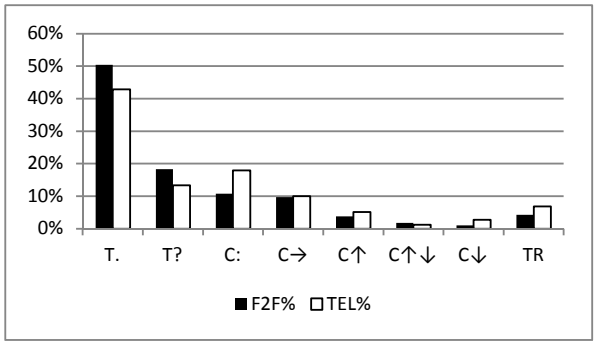

Figure 3: *Distribution of prosodic boundary tones in F2F and TEL.*

Statistically significance difference was found in the channel comparison, where the mean size values of several terminal units were found relatively much more frequent in TEL than in F2F. For example, 1-word T?-unit ($p=0.05$) and 3-words T?-unit ($p<0.01$) were found significant (among others); also 1-word T.-unit ($p<0.01$) and 2-words T.-unit ($p<0.05$). This can be explain as a discourse characteristic of TEL, where feedback are more frequent due to the lack of facial and other visual feedback. The question remains open for much larger units (8, 9, 10, 13 and 15 syllable units) that were found statistically significant. As to the T-size in terms of C-units, at the channel test this parameter was found significant in several cases. First, a difference was found in T-units with no C-units ($p<0.05$). Zero-C-units reflect mostly the short, feedback responses that were explained before as more common in TEL. It is interesting to see that also the T?-unit has some unique use in TEL: 1-C-boundary as well as 3, 8, and 9-C-boundaries were found significantly more frequent ($p<0.01$) in TEL. Perhaps this can be attributed to the opportunity to process longer stretches of speech in TEL, where the interlocutor is more attentive, with comparison to

more lively conversations in F2F dialogues. Significant was found also in the channel comparison of C-unit size. Each of the five C-units had at least one unit-size that demonstrated statistical significance. This correlates with the explanation of longer stretches of speech in TEL, which was mentioned before with regard to terminal units. Two more measurements were carried out on small sub-set of the corpus. First, since the 19 informants of each of the 19 recordings were the main speakers, a subset of Main speakers only was extracted. This subset consists of 8 men and 11 women. An Anova test with repeated measures was carried out, and no significant gender differences were found. For example, p=0.752 on T-unit size. This means that men and women use the same pattern of T-unit sizes (Large amount of 1-word T-units, much less 2-words T-units and declination in use till 15-words T-unit size). It seems that even when reducing the corpus to a less varied speech material, no gender differences are found. Another sub-corpus was extracted for channel pairs of the same speaker. The motivation behind this sub-corpus was to investigate if there is an intra-speaker channel difference. In this sub-corpus only eight speakers out of the 19 main speakers had both TEL and F2F speech material. A T-test was carried out and no significant channel difference was found (p>0.05). This means that speakers use the same patterns, with regard to IU size, in TEL and F2F conversations.

4.3. Syllable duration in various prosodic environments

In order to compare the durational parameter of the syllables in the five continuous boundaries, a pilot study of 22 minutes from TEL (female speaker) and 12 minutes from F2F (male speaker) were segmented manually into syllables. These two recordings were chosen since their acoustic quality meets basic acoustic measurement standards, such as a clear voice and an absence of background noises or speech overlaps. Moreover, in both recordings there is only one interlocutor (compared to other F2F recordings where normally consists of more than two interlocutors). Only the speech of the informant speaker was measured. In TEL, the number of IUs is 610; in F2F, 177. Figures 4 and 5 summarize the durational measurements taken. In this study, the duration of the entire syllable was measured. A threshold to the C: boundary tone was set on a minimum of 230ms. Therefore, the syllables that carry the C: tone were found in both recordings to have the highest mean values (rightmost B&W histograms in Figure 4), while the fluent syllables were found in both cases with the lowest mean values (leftmost B&W histograms in Figure 4).

4.4. Estimated IU size in IH

The estimated unit size was measured as a combination of two variables: 1. In order to avoid the "long tail" bias, we calculated the mean value (syllables) of the most frequent IU sizes; and 2. We used the durational measurements of syllables mentioned in subsection 4.3 above. The mean values of most frequent IU sizes are as follows (almost no difference between the two channels):

T. = 3 (TEL)-3.5 (F2F) syllables per IU
T? = 2.3 (TEL)-3 (F2F) syllables per IU
C→ = 5 syllables per IU
C: = 2.5 syllables per IU
C↑ = 4.5 syllables per IU
C↑↓ = 2 syllables per IU
C↓ = 6.6 syllables per IU

It is demonstrated in the above that the mean C-unit is *longer* than the average T-unit. This is exactly because the calculations took the frequency of use into consideration. Since above 75% of T-units are with no internal C-units, their mean size seems shorter. Second, it should be stressed here that due to the limited size of durational measurements, and to the mixed variables recordings, the results shown on Figures 4 and 5 are within the realm of estimation only. Again, the estimated duration of the two T-units does not include duration of C-units, since above 75% of T-units were without internal C-units.

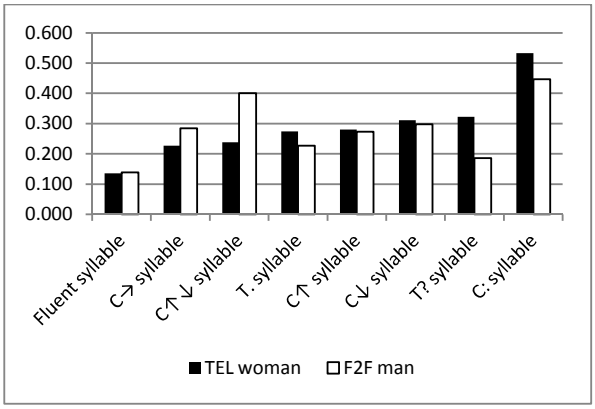

Figure 4: *Syllable duration (seconds) in TEL (woman) and in F2F (man).*

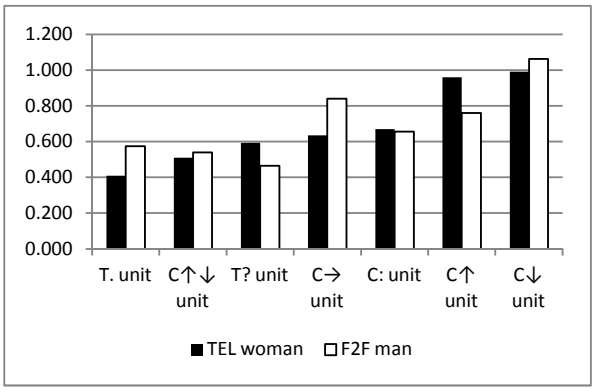

Figure 5: *Estimated IU size (seconds) in TEL (woman) and in F2F (man).*

5. Discussion

This preliminary study highlights the need to examine prosodic unit size in spontaneous IH. Prosodic unit size was examined with regard to syllable and word structures, and sub-units (C-units) in higher prosodic units in the hierarchy (T-units). TEL conversations are characterized by *single-word coherent contour* (T.)-units, while (T?)-units mostly consist of several IUs (with at least two C-units). The syllable durational measurements suggest that in IH, as in AE, "filled pauses differ dramatically from (…) other instances in duration" [11]. With regard to gender and channel differences, in both variables the use of T-boundaries versus C-boundaries was found to be statistically significant. Nevertheless, the duration measurements were carried out on a relatively small portion of the corpus, and syllable duration in spoken IH still needs to be investigated in future research.

6. References

[1] Schmid, S., “Phonological typology, rhythm types and the phonetics-phonology interface. A methodological overview and three case studies on Italo-Romance dialects”, in A. Ender, A. Leemann and B. Wälchli, (Eds), Methods in contemporary linguistics, 45-68, Berlin: Mouton de Gruyter, 2012.

[2] Lacheret, A., Bordal, G., and Truong, A. “Ch. 11: The prosodic structure”, in A. Lacheret, S. Kahane, and P. Pietrandrea, (Eds.). Rhapsodie: a prosodic syntactic treebank of spoken French. Amsterdam-Philadelphia: John Benjamins Publishing Company, 2014.

[3] Finkelstein, M. and Amir, O., “Speaking Rate among Professional Radio Newscasters: Hebrew Speakers”, Studies in Media and Communication 1(1):131-139, 2013.

[4] Hirose, K., and Kawanami, H., “Temporal rate change of dialogue speech in prosodic units as compared to read speech”. Speech Communication, 36(1):97-111, 2002.

[5] Jacewicz, E., Fox, R. A., O'Neill, C., and Salmons, J., “Articulation rate across dialect, age and gender”, Language Variation and Change, 21:233-256, 2009.

[6] Amir, O. and Grinfeld, D., “Articulation rate in childhood and adolescence: Hebrew speakers”, Language and Speech, 54(2):225-240, 2011.

[7] Silber-Varod, V., The SpeeCHain Perspective: Form and Function of Prosodic Boundary Tones in Spontaneous Spoken Hebrew, LAP LAMBERT Academic Publishing, 2013.

[8] Izre'el S. and Mettouchi, A., “Representation of Speech in CorpAfroAs: Transcriptional Strategies and Prosodic Units”, in A. Mettouchi, M. Vanhove, and D. Caubet (Eds), Corpus-based Studies of Lesser-described Languages: the CorpAfroAs Corpus od Spoken AfroAsiatic Languages. Amsterdam-Philadelphia: John Benjamins, to appear.

[9] COSIH: The Corpus of Spoken Israeli Hebrew <http://humanities.tau.ac.il/~cosih/english/>

[10] Segal, O., Nir-Sagiv, B., Kishon-Rabin, L., and Ravid, D., “Prosodic patterns in Hebrew child directed speech”, Journal of Child Language, 36(3):629-656, 2009.

[11] Shriberg, E. “To errrr' is human: ecology and acoustics of speech disfluencies”, Journal of the International Phonetic Association 31(1):153-169, 2001.