Speaker-specific intonational marking of narrow focus in Egyptian Arabic

Francesco Cangemi[1], *Dina El Zarka*[2], *Simon Wehrle*[1], *Stefan Baumann*[1], *Martine Grice*[1]

[1] I*f*L Phonetik, University of Cologne, Germany
[2] University of Graz, Austria
{fcangemi, swehrle2, stefan.baumann, martine.grice}@uni-koeln.de, dina.elzarka@uni-graz.at

Abstract

Experimental evidence suggests that prosodic encoding of information structure in Egyptian Arabic (EA) might be limited to contrastive focus and achieved through phonetically continuous means (i.e. pitch range expansion on the focused constituent and pitch compression on post-focal material). In this paper we explore the hypothesis of a richer mapping between information structure and intonation in EA, with respect both to the encoding of further focus types and to the use of parameters beyond pitch range (e.g. alignment of f0 turning points).

By performing speaker-specific analyses on a dataset of read speech from 18 speakers, we provide evidence suggesting that EA also uses intonation to encode narrow information focus. For many speakers this is achieved through a different temporal alignment of f0 turning points instead of, or in addition to, pitch range manipulation.

Crucially, the findings highlight the usefulness of a speaker-specific analysis for the study of the mapping between prosody and information structure.

Index Terms: speaker-specificity, intonation, information structure, Egyptian Arabic, alignment

1. Introduction

1.1. Egyptian Arabic intonation

Egyptian Arabic (EA) has featured prominently in prosodic typology research as an intonation language with a number of distinct aspects [1]. According to research in the autosegmental-metrical framework [2], the intonational inventory of EA is limited to a single pitch accent type [3]. This might be related to the fact that EA, unlike languages with larger pitch accent type inventories [4], does not appear to use prosody to encode information status (i.e. givenness or newness) [5].

Along these lines, the use of prosody in EA also seems to be limited in its encoding of other levels of information structure, in two respects at least. First, specific types of focus are encoded using continuous pitch range adjustments only. For example, contrastively focused constituents are only flagged through pitch range expansion on the focused constituent and through pitch range compression on postfocal material [6]. Second, intonation is used to encode fewer information structural contrasts than in other languages. For example, no prosodic reflexes of information focus have been documented for EA [7].

Further research suggests that EA prosody might be more sensitive to information structure [8], both in terms of using parameters beyond pitch range adjustments and in terms of the number of contrasts encoded through prosody. According to this view, topic constituents display rises or rise-plateaux, whereas focal constituents display (rise-)falls. Moreover, preliminary evidence [9] suggests that the contrast between narrow information focus and broad focus might be encoded prosodically too. The interpretation of the findings is however obfuscated by the presence of a wide range of speaker-specific strategies in the prosodic encoding of the two focus types.

1.2. Speaker-specificity

Unlike research in the segmental domain, intonational phonology has only recently started to focus on individual-specific strategies (see [10] for a brief overview). A number of studies on German ([11] and [12] for production, and [10] for production, perception and their interaction) suggests that individual speakers might use different sets of cues to encode the same intonational contrast. [11] show that speakers can rely more heavily on either peak shape or peak alignment in the encoding of the contrast between given and new material (in terms of argumentation structure, see [13]). Furthermore, [12] show that speakers can use a variety of intonational and supralaryngeal devices in encoding the contrast between broad, narrow and contrastive focus.

We argue that a targeted investigation of speaker-specific strategies is of even greater importance if we want to determine whether a given contrast is encoded prosodically *at all*, as in the case of the controversy on EA narrow information focus mentioned above. Crucially, both the studies suggesting limited [5][6][7] and extended [8][9] sensitivity of EA intonation to information structure found massive speaker-specific effects, even with a relatively low number of speakers (only 6 in all of the studies cited above). In this paper, we thus explore prosodic encoding of narrow information focus in EA through the prism of speaker-specific strategies.

1.3. Rationale

By combining the hypotheses above on the use of f0 falls for focused constituents [8] and on the prosodic encoding of narrow focus [9] in EA, we predict that f0 falls will have different properties in all-focus and narrow information focus utterances with the same lexical material.

We operationalize differences in f0 falls by measuring scaling and alignment (with respect to the end of the stressed syllable) of its start and end points, defined as the high and low turning points in the f0 contour, respectively. For comments on alternatives to this operationalization choice, see §4.2. No specific predictions on scaling or alignment values are made. However, since narrow information focus displays intermediate characteristics between all-focus and contrastive focus in a number of languages, and since contrastive focus is assumed to be encoded with expanded pitch range in EA, the scaling of high turning points could be expected to be higher in narrow information focus than in broad focus condition.

The two measures (scaling and alignment) for the two turning points (high and low) are employed in both a traditional across-speakers global analysis and a novel speaker-specific profiling analysis. The speaker-specific analysis is expected to both validate the findings from the global analysis and to provide evidence for further encoding strategies that cannot be profiled using pooled data.

2. Method

2.1. Participants, procedure and stimuli

Productions were recorded from 18 native speakers of EA (11 f). Most of them were either in their twenties or in their fifties, were born and raised and living in Cairo or Alexandria, had university-level education, studied English and/or French but did not use these languages regularly in their private life, and had EA-speaking partners. Recordings took place in Cairo and Alexandria.

Stimuli were composed of 6 experimental sentences and 4 filler sentences in Arabic script presented on a computer screen. Participants read the sentence after listening to a pre-recorded question stimulus that was preceded by a contextualizing paragraph. The different questions and contexts were meant to elicit a variety of information structure conditions for the experimental sentences. In this study, we concentrate on the all-focus or broad focus condition (henceforth BF) and on the narrow information focus condition (NF). The examples below provide translations of a target sentence (2) as an answer to a BF-inducing context (1a) and a NF-inducing context (1b):

(1a) ħasˤal ʔe: min waʔtə ma-na safirt?
What has happened since I left the country?

(1b) bila ʃakk, il-luɣa:t il-ʔagnabijja baʔit muhimma ʔawi l-jome:n do:l, fi: ħaddə tʕallim ʔalma:ni?
No doubt, foreign languages are very important nowadays, has anyone learned German?

(2) Sali:ma tʕallimit ʔalma:ni.
Salima learned German.

Target sentences were composed by target word (a proper name in subject position) followed by a verb and an object. Target words were trisyllabic and had penultimate stress, the stressed syllable containing the long vowel /i:/ surrounded by sonorants or /h/.

Speakers read each target sentence three times in all conditions, for a total of 18 speakers * 6 sentences * 2 conditions * 3 repetitions = 648 recorded items. 58 items (~9%, of which 14 for speaker M02 and 18 for speaker F06, see §4.2) had to be excluded because of noticeable disfluencies, major f0 discontinuities that made turning point detection meaningless, and utterances which differed from the target sentence. After excluding further 7 items due to the quality of the recordings, 583 items underwent analysis (~90%).

2.2. Measures

The second author manually segmented target words into syllables using *Praat* [14]. F0 tracks for all items were extracted, manually corrected and smoothed by the third author using the *Praat* script *mausmooth* [15]. The smoothed contours were used for automatic extraction of the f0 fall associated with the target word. High turning points (fall startpoint) were defined as f0 maxima in a window going from the beginning of the item to 150ms after the end of the stressed syllable in the target word. Low turning points (fall endpoint) were detected automatically using R [16] as second derivatives maxima in a window spanning from the high turning point to 300ms after the end of the stressed syllable. The right boundaries of the analysis windows were set after visual examination of superposed f0 contours for the entire dataset (also produced with *mausmooth*).

The results of the automatic detection were manually verified by the third author. All high turning points except one were correctly placed, and 62 low turning points had to be moved. This mostly happened when the analysis window was large enough to have the second derivative maximum capture the beginning of a subsequent f0 rise, rather than the endpoint of the f0 fall on the target word. In 39 of these cases, in order to reduce arbitrariness of manual repositioning, low turning points were located using two-line regression (for both methods see [17, pp. 44-46] and references therein). The elbows were used to characterize the f0 fall in terms of scaling (in Hz) and alignment (in ms, relative to the end of the stressed syllable).

Figure 1 shows an example of a test utterance for the sentence in (2) uttered as NF with its waveform (in pink), original f0 track as extracted by *Praat* (grey speckles), corrected f0 track using *mausmooth* (black solid line) in the top panel, broad phonetic transcription and syllable segmentation for the target word (first tier, stressed syllable capitalized) and automatically extracted turning points (second tier) in the bottom panel. The thick lines in the top panel show the extracted measures: alignment (dotted) and scaling (solid), for the high (black) and low (grey) turning points.

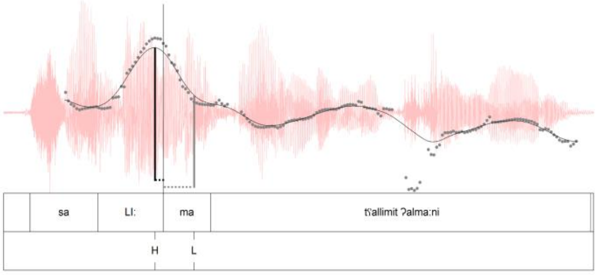

Figure 1: *Example of segmentation, smoothing, elbow location and extracted measures.*

3. Results

3.1. Global analysis

Four linear mixed effects models were used to predict alignment and scaling of high and low turning points {*H_alig, H_scal, L_alig, L_scal*} with Type {*BF, NF*}, Repetition {*X,…Z*} and their interaction as fixed effect, and with random slopes and intercepts for Target sentence {*A,…F*} and Speaker {*F01,…M07*} [18]. In all cases, Likelihood Ratio Tests showed that the interaction between Type and Repetition could be dropped without loss of fit. Removing Repetition yielded instead fit loss for *H_alig* only; a closer examination showed however that (i) Repetition coefficients reached significance (*t*=-3.1) only for the third repetition, (ii) Repetition coefficient size is smaller than Type coefficient size, and (iii) Repetition coefficient sign is always same as Type coefficient sign. We interpret this pattern of results as suggesting that any documented Type effect increases across repetitions, but is already attested from the first instance. As a consequence, in the following we evaluate TYPE in the minimal adequate models (which conservatively include for all dependent variables TYPE and REPETITION but not their interaction) and do not further discuss REPETITION effects.

Table 1. *Models estimates and test statistics.*

Variable	Int (BF)	TYPE (NF)	χ^2	p
H_alig	27	-31.5	85.7	<0.01
H_scal	214.5	11.9	51	<0.01
L_alig	228.2	-57	103	<0.01
L_scal	171.4	-1	0.7	0.4

The significance of TYPE was assessed for each model through Likelihood Ratio Tests against a null model without TYPE. For each dependent variable, Table 1 shows estimates for intercept (BF) and for TYPE (NF), and χ^2 and p-values from LRTs. The results show that, with the exception of *L_scal*, all comparisons were significant at α=0.05. Both turning points were aligned earlier (negative TYPE coefficients) and high turning points were scaled higher (positive TYPE coefficient) in the NF condition. Note that the duration of the target syllables did not differ across the two conditions. Figure 2 plots TYPE standard errors and estimates from the four models for intercepts (BF, black circles) and TYPE (NF, red squares) for high (filled symbols) and low (empty symbols) turning points [19].

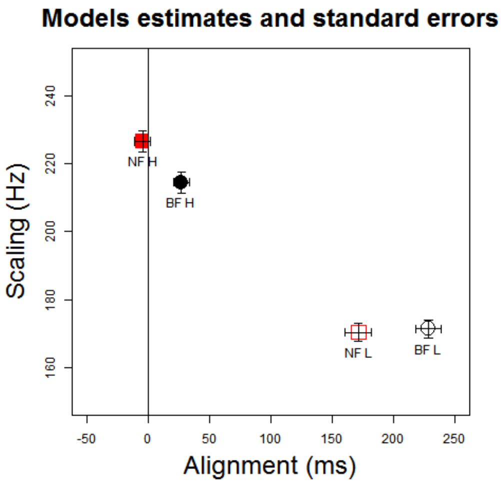

Figure 2: *Global analysis: estimated means for intercept (BF) and Type (NF) with standard errors for Type. The vertical line indicates the end of the stressed syllable.*

An informal examination of results from individual speakers shows indeed that, unlike alignment, scaling is subject to speaker-specific strategies, in particular for the low turning point. This is shown in Figure 3, where alignment (x-axis) and scaling (y-axis) are plotted for actual data points for two speakers (left panel: F03, right panel: F02), using filled and empty symbols for high and low turning points respectively, and black circles and red squares for BF and NF respectively. Whereas alignment of both turning points is consistently earlier in NF condition for both speakers, scaling of both turning points is higher only for speaker F03. For speaker F02, high turning points show comparable scaling in the two conditions, while low turning points are actually scaled lower in NF.

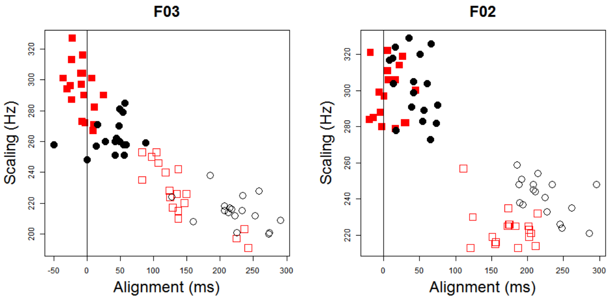

Figure 3: *Observed values of alignment (x-axis) and scaling (y-axis) of high (filled) and low (empty) turning points in BF (black circles) and NF (red squares) condition for two speakers.*

3.2. Speaker-specific analysis

In order to assess speaker-specific strategies in the alignment and scaling of high and low turning points as phonetic exponents of focus types, we ran a series of linear mixed effect models with the same structure as the minimal adequate models from the global analysis, thus predicting alignment and scaling of high and low turning points $\{H\_alig, H\_scal, L\_alig, L\_scal\}$ with TYPE $\{BF, NF\}$ and REPETITION $\{X, \ldots Z\}$ as fixed effect, and with random slopes and intercepts for TARGET SENTENCE $\{A, \ldots F\}$. Each model was fitted to data from a single speaker at a time. The significance of TYPE was assessed for each model through Likelihood Ratio Tests against a null model without TYPE.

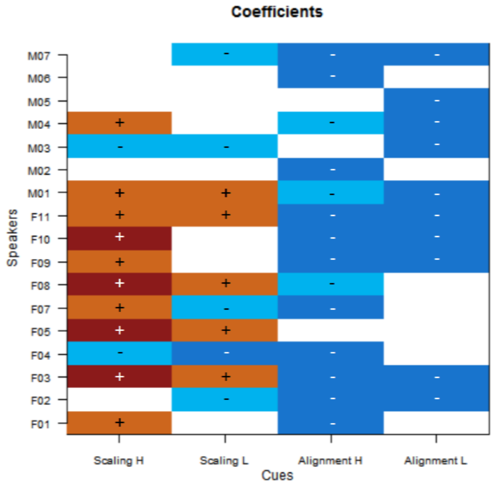

Figure 4: *Speaker-specific analysis: coefficients for individual cues (x-axis) and speakers (y-axis); white cells are used for non-significant LRTs; red and blue cells for positive and negative coefficients; orange and cyan cells for negligible effects.*

Figure 4 shows results for each cue (x-axis) and for each speaker (y-axis). Cells are left white if LRTs do not reach significance; for results reaching significance, cells are filled in red or orange (with a + sign) if the coefficients are positive (indicating higher scaling or later alignment) and are filled in blue or cyan (with a - sign) if the coefficients are negative (indicating lower scaling or earlier alignment) [20]. Lighter shades (orange and cyan, for positive and negative coefficients, respectively, using black + and - signs) are used if LRTs reach significance but the coefficient size is small, suggesting perceptually negligible effects. The cutoff point was set at 20ms for alignment differences and at 10% of the intercept value for scaling differences. For example, for speaker F08 the LRT between the full and null models predicting alignment of the low turning point does not reach significance, and the corresponding cell is thus left white. The LRT between the full and null models predicting alignment of the high turning point reaches significance ($\chi^2(1)=8.1$, $p<0.01$), and the full model has a coefficient for TYPE of -13.7, thus below the 20ms threshold for strong effects; the corresponding cell is thus color-coded as cyan (significant, negative and negligible). The LRT between the full and null models predicting scaling of the high turning point reaches significance ($\chi^2(1)=41.7$, $p<0.01$), and the full model has a coefficient for TYPE of 24.4; since the intercept has a coefficient of 211.4, the effect is above the 10% threshold for strong effects; the corresponding cell is thus color-coded as red (significant, positive and strong).

The speaker-specific analysis corroborates the results of the global analysis: both high and low turning points are generally aligned earlier in NF condition, as attested by the predominance of significant negative coefficients (blue cells in the two rightmost columns). High turning points are generally scaled higher in NF condition, as attested by the predominance of significant positive coefficient (red cells in the leftmost column). For scaling of low turning points (second column), the results are less clear-cut, mirroring the negligible coefficient size of the global analysis. Moreover, the speaker-specific analysis highlights a number of further interesting findings.

First, no single cue is consistently used by all speakers, as attested by the absence of columns entirely filled with dark cells. Second, while some cues are used rather consistently by speakers (as high turning point alignment, with 14 significant coefficients, 11 of which of large size, and all with the same sign), other cues are more prone to speaker-specific strategies (as low turning point scaling, with 10 significant coefficients, only 1 of which of large size, and equally split in the two signs). Third, no single speaker uses all cues consistently. Even speakers whose contrasts reach significance for all cues only show strong effects in three (F03), two (F11) or one (M01) case out of four. Fourth, while speakers vary in the number and strength of cues they employ, all of them have at least one large-size coefficient across the four cues (M05).

Taken together, our findings converge towards the existence of speaker-specific strategies in the intonational encoding of focus type in EA.

4. Discussion

4.1. Intonation and information structure in EA

Our global and speaker-specific analyses support a heightened extended sensitivity of EA intonation to information structure, consistent with the findings in [8][9]. This applies both to the prosodic encoding of more than one contrast and to the use of parameters other than pitch range in such encoding. As for the first point, the speaker-specific analysis shows that, even if through the use of different cues, the contrast between all-focus and narrow information focus is indeed encoded prosodically by all speakers. This finding suggests that, by using an appropriate methodology, further information structural contrasts might be documented for EA. As for the use of parameters beyond pitch range in the encoding of focus type, both the global and the speaker-specific analyses converge in showing that alignment differences between the two focus conditions are stronger than scaling differences, especially for low turning points.

This is not to say that scaling does not play a role in EA. Actually, our data show that scaling of the high turning point is almost consistently higher in narrow information focus condition, while scaling of the low turning point is higher for some speakers and lower for others. These results are compatible with the two continuous pitch modifications acknowledged in the literature, namely pitch register vs. pitch range [21] or pitch level vs. pitch span [2], for raised vs. lowered low turning points, respectively. Incidentally, both pitch register/level and pitch range/span modifications are also consistent with the stable higher scaling of high turning points. Crucially, speakers are able to make use of either of the two modification types available.

4.2. Theoretical implications and further research

The findings above raise further questions for the study of EA intonation and for phonological theories of intonation in general. First, the phonetic differences between the two focus conditions were operationalized here in terms of scaling and alignment of the beginning and end of the f0 fall on the focused constituent. We are confident that other cues might be at play here, and that many more columns should be added to the heatmap in Figure 4. For example, the initial rising f0 movement leading to the beginning of the fall can also be expected to have different properties in the two conditions. Moreover, in §2.1 we mentioned that data for low turning point scaling and alignment was not reliable for speaker F06, due to problems in the elbow location procedure stemming from the presence of a silence after the target word in all 18 narrow focus items (for this reason, no coefficients are plotted for speaker F06 in Figure 4; the 14 items excluded for speaker M02, also mentioned at §2.1, were instead evenly split across BF and NF conditions, thus allowing statistical analyses). It is clear that the insertion of a prosodic break after the focused constituent might be yet another strategy available to speakers to encode focus types.

Second, it will be interesting to explore whether some of the strategies available to the speakers are actually equivalent in perception. Different patterns in the alignment and scaling of individual turning points might be decoded as holistic percepts. Integrated measures such as the tonal center of gravity proposed by [22] might be used to explore the perceptual equivalence of different encoding strategies.

Third, having documented intonational encoding of narrow information focus by EA *speakers*, it will be interesting to see to what extent EA *listeners* use such cues in decoding focus types, ideally using a variety of speech styles. Reliable decoding would then raise the issue of an appropriate phonological modeling of intonational devices for the expression of focus types, e.g. (for autosegmental-metrical accounts) in terms of pitch accent inventories. This would, in turn, lead to a refinement of the currently available prosodic typologies, in which EA represents an instance of a single-pitch-accent language.

Lastly, as suggested by [10] the study of the multiple paths to the encoding of information structural contrasts is crucial to a refined understanding of intonational categories – and of phonological categories *tout court*.

5. Acknowledgements

The first author's work is funded by the German Research Foundation Excellence Initiative at the University of Cologne (Emerging Group: Dynamic Structuring in Language and Communication). This collaboration stems from the first author's Guest Lectureship at the Signal Processing and Speech Communication Laboratory (Technical University Graz), promoted by Dr. Barbara Schuppler.

6. References

[1] S.A. Jun, "Prosodic typology: by prominence type, word prosody, and macro-rhythm", in *Prosodic Typology II. The Phonology of Intonation and Phrasing,* S.-A. Jun, Ed. Oxford: Oxford Univ. Press, 2014, pp. 520-539.

[2] D.R. Ladd, *Intonational Phonology*, Cambridge: Cambridge University Press, 2008.

[3] D. Chahal and S. Hellmuth, "Comparing the Intonational Phonology of Lebanese and Egyptian Arabic", *Prosodic Typology, 2,* pp. 365-404, 2014.

[4] M. Grice, S. Baumann and R. Benzmüller, "German Intonation in Autosegmental-Metrical Phonology", in *Prosodic Typology II. The Phonology of Intonation and Phrasing*, S.-A. Jun, Ed. Oxford: Oxford Univ. Press, 2014, pp. 55-83.

[5] S. Hellmuth, "No de-Accenting in (or of) phrases: evidence from Arabic for cross-linguistic and cross-dialectal prosodic variation", in *Prosodies: With Special Reference to Iberian Languages*, S. Frota et al., Eds. Berlin: Mouton de Gruyter, 2005, pp. 99-121.

[6] S. Hellmuth, "Acoustic cues to focus and givenness in Egyptian Arabic", in *Instrumental Studies in Arabic Phonetics*, Z. M. Hassan & B. Heselwood, Eds.Amsterdam: Benjamins, 2011, pp. 301-323.

[7] S. Hellmuth , "Focus-related pitch range manipulation (and peak alignment effects) in Egyptian Arabic", in *Speech Prosody 2006,* Dresden, Germany, pp. 410-413, 2006.

[8] D. El Zarka, "Leading, linking, and closing tones and tunes in Egyptian Arabic: What a simple intonation system tells us about the nature of intonation", in *Perspectives on Arabic Linguistics,* XXII-XXIII, E. Broselow and H. Ouali, Eds. Amsterdam: Benjamins, 2011, pp.57-73.

[9] D. El Zarka, "On the Interaction of Information Structure and Prosody: the Case of Egyptian Arabic*",* habilitation thesis, Dept. of Linguistics, Karl-Franzens-Universität Graz, 2013.

[10] F. Cangemi, M. Krüger and M. Grice, "Listener-specific perception of speaker-specific production in intonation", in *Individual Differences in Speech Production and Perception*, S. Fuchs et al., Eds. Frankfurt: Peter Lang, 2015, pp. 123-145.

[11] O. Niebuhr et al., "Are there 'shapers' and 'aligners'? Individual differences in signalling pitch accent category" in *17th ICPhS*, Hong Kong, China, pp.120-123, 2011.

[12] D. Mücke and M. Grice, "The effect of focus marking on supralaryngeal articulation – Is it mediated by accentuation?", *Journal of Phonetics, 44,* pp. 47-61, 2014.

[13] O. Niebuhr, "The signalling of German rising-falling intonation categories – The interplay of synchronization, shape and height, *Phonetica, 64,* pp.174-193, 2007.

[14] P. Boersma and D. Weenink (2015). *Praat: doing phonetics by computer (version 6.0.05)* [Online]. Available: http://www.praat.org/

[15] F. Cangemi (2015). *mausmooth* [Online]. Available: http://phonetik.phil-fak.uni-koeln.de/fcangemi.html.

[16] R Core Team (2015). *R: a language and environment for statistical computing* [Online]. R Foundation for Statistical Computing, Vienna, Austria. Available: https://www.R-project.org

[17] F. Cangemi, *Prosodic detail in Neapolitan Italian,* Berlin: Language Science Press, 2014.

[18] D. Bates et al., "Fitting Linear Mixed-Effects Models Using lme4", *Journal of Statistical Software, 67*(1), pp. 1-48, 2015.

[19] W. Revelle (2015). *psych: Procedures for Personality and Psychological Research* [Online]. Northwestern University, Evanston, Illinos. Available: http://CRAN.R-project.org/package=psych Version=1.5.8.

[20] A. Gelman and Y.-S. Su (2015). *arm: Data Analysis Using Regression and Multilevel/Hierarchical Models* [Online]. R package version 1.8-6. Available: http://CRAN.R-project.org/package=arm.

[21] C. Gussenhoven, *The phonology of tone and intonation,* Cambridge: Cambridge University Press, 2004.

[22] J. Barnes et al., "Tonal Center of Gravity: A global approach to tonal implementation in a level-based intonational phonology", *Laboratory Phonology, 3*(2), pp. 337-383, 2012.