



Neutral Tone Variation in Beijing Mandarin: Is Neutral Tone Toneless?

Xiao Dong¹, Fengming Liu², Chien-Jer Lin¹, Monica Nesbitt³, Shuju Shi³

¹Department of East Asian Languages and Cultures, Indiana University Bloomington, USA

²Department of Second Language Studies, Indiana University Bloomington, USA

³Department of Linguistics, Indiana University Bloomington, USA

{dong1, fl17, chiclin, nesbittm, shi16}@iu.edu

Abstract

Neutral tone (NT) is a distinctive feature of Beijing Mandarin, traditionally described as toneless and entirely dependent on the preceding tone. Recent studies suggest that NT may retain specific phonetic targets and exists on a continuum of reduction, challenging the strict neutral versus full-tone dichotomy. This study examines the phonetic realization of NT as influenced by three factors—preceding tone, underlying tone, and NT type—using word-list reading data from 36 Beijing Mandarin speakers. Our findings confirm a robust effect of the preceding tone. In the meantime, we identify a significant impact of the underlying tone, indicating that NT is not entirely toneless but retains some degree of phonological specification. Moreover, the differences observed between optional and forbidden NT words suggest that NT should be conceptualized as part of a gradient system influenced by contextual factors, rather than as a simple neutral versus full-tone contrast.

Index Terms: neutral tone, Mandarin

1. Introduction

Mandarin Chinese is a tonal language, where pitch plays a crucial role in distinguishing meaning. In addition to its four lexically distinctive full tones, Tone 1 (T1, high level 55), Tone 2 (T2, rising 35), Tone 3 (T3, low dipping 213), and Tone 4 (T4, falling 51), Mandarin features neutral tone (NT), which is considered as the 5th tone (T0) and usually occurs on non-initial unstressed syllables [1, 2, 3].

Previous studies have examined the acoustic characteristics of NT syllables. A Key finding is that NT syllables generally have shorter duration than fully stressed syllables, with some studies reporting their length as approximately 45% of their non-neutral counterparts (eg., [4, 1, 2]), while others suggest a ratio closer to 60% (eg., [5, 6]). Additionally, the pitch contour of NT syllables is influenced by the preceding stressed syllable's tone, although its precise realization varies across studies. For example, [4] found that for male speakers, NT was realized as 41, 51, 44, 42 respectively after T1, T2, T3, and T4. And for female speakers, the realizations are 41, 51, 33/32, and 21 respectively. [5] found the NT is realized as mid-level tone after T3 and mid-falling after the other three tones. [1] described mid-falling after T1, high-falling after T2, mid-level after the variant 21 of T3, and low-level after T4.

However, most of the previous studies have primarily focused on contrastive NT words (e.g., *dong1xi1* 'east and west' vs. *dong1xi0* 'stuff'), suffixes (e.g., *-zi*), reduplicated nouns (e.g., *ma1ma0* 'mom'), and other words where NT is obligatory [4, 2, 6]. More recent studies have started incorporating different types of neutral syllables in their studies, and suggest that these types of neutral syllables vary in degree of reduction

rather than fitting into a strict "neutral vs. full" dichotomy. For example, [7] investigated the acoustic properties of nine types of NT words, including structural particles (e.g., *-de*), aspect markers (e.g., *-zhe*), suffixes, locative words (e.g., *-shang*), directional verbs (e.g., *-lai*), reduplicated nouns, habitual NT words, optionally neutralized words, and functional NT words. Their findings revealed systematic variation across these categories: structural particles, suffixes, aspect markers, and functional NT words exhibited the greatest degree of reduction, while reduplicated words, directional verbs, and locative words formed an intermediate category. Optional NT words emerged as a distinct category and showed considerable variability. However, [7] did not examine the variations within optional NT words, leaving unanswered questions about how these words are realized differently and what factors influence their realization.

Another overlooked category in NT research is forbidden NT words. The Beijing dialect exhibits more NT words than Standard Mandarin [8]. Forbidden NT words refer to words that are neutralized in the Beijing dialect but are expected to retain full tone in Standard Mandarin. For Standard Mandarin speakers, these words are expected to be pronounced with full lexical tone. For Beijing speakers, they may be variably neutralized, but how they behave acoustically remains unknown. If NT operates along a gradient, as [7] suggested, then forbidden NT words may fall into a less-reduced category, showing distinct duration, pitch, and intensity compared to more fully reduced NT words. To better understand these patterns, the current study investigates the acoustic characteristics of these two understudied types of NT.

In addition, due to the strong contextual dependency of the NT contour, it has been widely accepted that NT has lost its original tonal identity and is inherently toneless, lacking an independent underlying pitch target [9, 10, 5, 11]. Some different views include [12] and [6]. [12] regards NT as a toneless element at the phonological level but not as register-less. The author proposes that in the underlying representation, NT in Mandarin is specified with '-Upper' in the Register tier but unspecified in the Tone tier. The specification of NT on the Tone tier comes from target spreading of the preceding tone. In other words, the surface realization then is a consequence of the interplay between '-Upper' and the target from the preceding tone. More recently, [6] proposed that NT has a mid-low target of its own regardless of its surrounding contexts, and the relatively large amount of variation observed is attributed to the ineffectiveness of NT in overcoming the influence of the preceding tone as the weak element in its realization.

Despite these debates, previous studies generally agree that the underlying tone of a NT syllable does not directly affect its realization. This raises the question of whether this also applies to NT types that exist on a more lexical continuum, such

as the optional and forbidden NT discussed above, which may retain more characteristics of full tones. Do these NT syllables remain truly toneless, deriving their contour entirely from the preceding tone? Or do they exhibit a single target tone or register specification? Alternatively, do they show influences from their underlying tones?

If these NT syllables are truly toneless, we would expect no systematic influence from factors beyond the preceding tone. Conversely, if they have a fixed phonetic target, their F0 contours should consistently converge toward that target, irrespective of linguistic conditions. On the other hand, they may possess multiple targets, exhibiting systematic variation influenced by specific linguistic factors.

To answer these questions, we collected word list reading data from 36 Beijing Mandarin speakers. The main questions we ask are:

1) What are the main acoustic characteristics of optional and forbidden NT words?

2) How do the preceding tone and the underlying tone of NT syllables influence the acoustic properties of these two NT types?

By answering the first question, we aim to provide an empirical description of the realization of these two understudied NT types. This allows us to test whether different NT types exhibit distinct features and exist along a continuum. The second question offers insights into the nature of these two NT types, specifically whether they are truly toneless.

2. Data

We collected word list reading data from 36 Beijing speakers (20 female, 16 male), aged 25 to 45 years. Our goal was to include 48 optional and 48 forbidden NT words as stimuli. Within each group, there would be 16 nouns, 16 verbs, and 16 adjectives, ideally covering all 16 possible tonal combinations. However, among the forbidden NT words, we were unable to identify verbs with T1+T3 or T3+T3 tonal combinations, nor adjectives with T1+T4 or T3+T3 combinations. As a result, 48 words were included in the optional group and 44 in the forbidden group.

To balance the word list and minimize participants' attention to neutral tone, we also added 96 non-neutral tone filler words. The words were presented on a laptop screen using a PDF document. Each speaker read each word once, except for five participants, who read each word twice. For these speakers, we retained only their second pronunciation to minimize coarticulation effects. All recordings were collected at a sampling rate of 44,100 Hz.

2.1. Data segmentation

Since tones in Mandarin are primarily carried by the finals, we begin by segmenting each syllable into their initials and finals. This segmentation is semi-automated: we use the Montreal Forced Aligner to train acoustic models on our speech data and then perform forced alignment on the same dataset [13]. The forced-aligned results are subsequently manually checked and adjusted by two well-trained linguists using Praat ([14]).

2.2. Neutral tone annotation

Two native speakers of Mandarin Chinese manually annotated each token for every speaker, labeling a token with 0 when a NT was perceived and 1 otherwise. The results revealed that optional NT words were neutralized 42.5% of the time, while for-

bidden NT words were neutralized 27.5% of the time, with an inter-annotator agreement rate of 92%. Only neutralized words were included in the following analysis.

2.3. Acoustic measurement

This study focuses on two primary acoustic features: duration and F0. Duration was measured for both the neutral syllable and its preceding syllable to enable direct comparison. F0 values were extracted exclusively from the finals of the neutral syllable, sampled at ten equally spaced time points (5%, 15%, 25%, ..., 95%).

F0 values were extracted using the Python interface for Praat (Parselmouth) [14]. In the initial extraction, we applied standard settings: a pitch ceiling of 600 Hz and a pitch floor of 100 Hz for female speakers, and a pitch ceiling of 400 Hz with a pitch floor of 75 Hz for male speakers. Data inspection revealed that pitch-doubling errors were prevalent, likely due to pronounced creakiness in the latter half of the neutral syllable—particularly when the preceding syllable carried Tone 4, where creakiness sometimes spanned the entire syllable. While previous research on neutral tones typically excluded creaky data points, our approach attempted to tackle this challenge in two ways:

- After the initial extraction, we computed each speaker's mean F0 (in Hz). For syllables preceded by T1, T2, or T3, we re-extracted the final five F0 values using a pitch floor of 30 Hz and a pitch ceiling set to the speaker's mean F0 plus 100 Hz. In cases where the preceding tone is Tone 4, we extracted all ten F0 values using the creaky setting.
- Smoothing the F0 Contour: A three-point moving average was applied to the entire F0 contour to further reduce the influence of octave errors.

Further examination indicates that this approach enhances the accuracy of pitch extraction, particularly in handling creaky sounds.

We then normalized the F0 values using the following formula, which facilitates cross-speaker comparisons and aligns neutral tone contours with Yuanren Chao's five-tone scale [9][15]:

$$T = [(\lg X - \lg L) / \lg H - \lg L] * 5$$

3. Analysis

We conducted statistical analyses on both duration and F0 contours of NT using R [16]. For duration analysis, we first compared the duration of NT syllables with their preceding syllables using ANOVA [17]. Then we examined how the duration of NT syllables is influenced by its preceding tone, its underlying tone, and NT type, along with their interactions, using mixed-effects linear models [18]. Model comparison was conducted to determine the best-fitting model. Speaker and Word were included as random factors to account for individual variability and item-based differences.

For F0 modeling, we applied GAMMs using the bam function of the mgcv package [19] [20]. GAMMs were chosen for F0 modeling because they allow for the inclusion of a time variable, making it possible to analyze dynamic F0 patterns and track how F0 contours evolve over time. The dependent variable was the normalized F0 measured at ten time points. Given that previous research consistently show the contour of a NT is mostly influenced by the preceding tone, we modeled the effect of the preceding tones first. Within each preceding tone

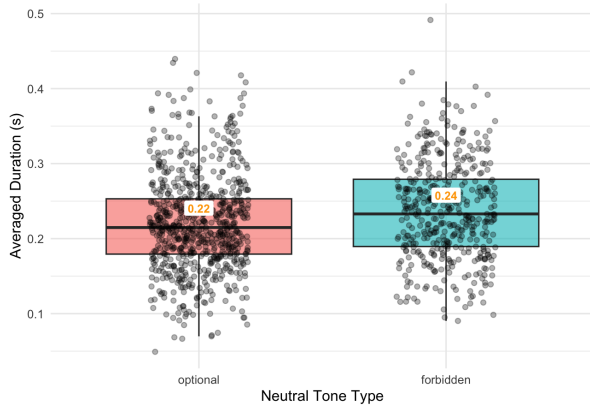


Figure 1: Averaged duration of neutral tone syllables by neutral tone type.

category, we further modeled the effects of the underlying tone of the NT syllable and NT type. This hierarchical structure enables the examination of how different underlying tones and NT types influence the F0 contour under each specific preceding tone context. To account for inter-speaker variability, we included a non-linear random effect for each speaker. Moreover, model autocorrelation was conducted to decrease temporal dependencies, and model criticism was conducted to identify potential areas of misfit, ensuring the robustness and reliability of the results.

4. Results

4.1. Duration

Our results show that NT syllables have significantly shorter duration than their preceding tones (averaged value: 0.23s vs. 0.3s, $p < 0.001$). In addition, optional NT words have significantly shorter duration than forbidden NT words ($p < 0.01$; see Figure 1). We also found a significant effect of preceding tone, where NT syllables preceded by T3 are significantly shorter than those preceded by T1 or T4 ($p = 0.01$ and $p = 0.03$ respectively). For the underlying tone of the NT syllable, T2 leads to significantly longer duration than T3 ($p < 0.0001$) or T4 ($p = 0.01$), and T1 results in significantly longer duration than T3 ($p < 0.0001$; see Figure 2).

In addition, the results show several significant interactions. The combinations T3 (preceding) + T2 (underlying), T3 + T3, and T4 + T3 significantly lower duration. Preceding T2 significantly lengthens duration under the neutral tone type (optional) condition.

4.2. F0 contour

Figure 3 illustrates the F0 contours for NT syllables following each of the four preceding tones. A visual comparison reveals that NT syllables generally exhibit a falling F0 contour after all preceding tones, except when following T3, where a rising trend is observed at the final time points. Additionally, NT after T2 tend to have a higher F0 than those following other tones, although this elevation is later surpassed by the rising contour of NTs after T3. In contrast, NTs after T1 and T4 show lower F0 contours.

To assess whether these F0 differences are statistically significant, we examined the difference curves. Figure 4 displays

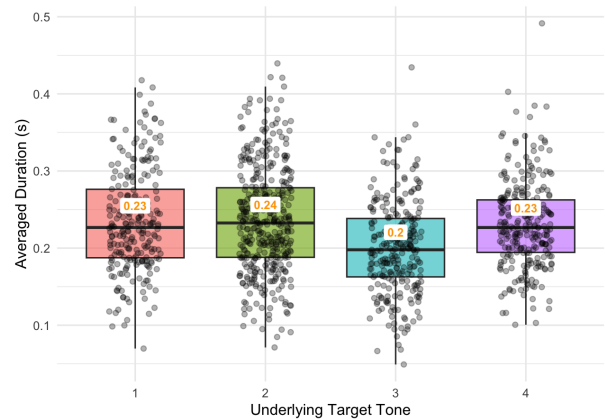


Figure 2: Averaged duration of neutral tone syllables by the underlying tone of neutral tone syllables.

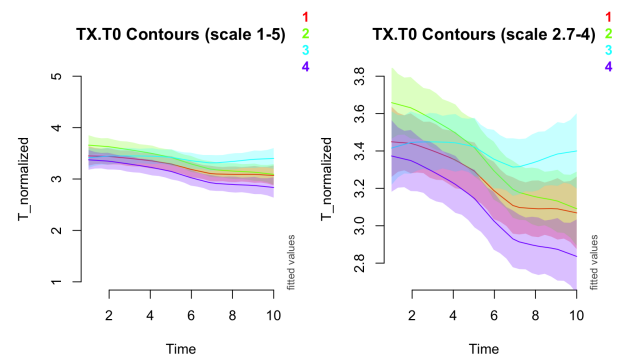


Figure 3: Neutral tone contours after each preceding tone.

the estimated difference curves for neutral tone contours between each pair of preceding tones. The 95% confidence interval is represented by a shaded band. Significant differences, where the confidence band does not overlap with the x-axis (i.e., the value is significantly different from zero), are indicated by red lines along the x-axis and vertical dotted lines.

The results show that the NT contour after T4 is significantly lower than that after T1, T2, and T3. Additionally, the NT contour following T1 is lower than that following T2 during the first seven time points and lower than that following T3 at the last six time points. Moreover, NT contour following T2 is higher than that after T3 during the first four time points but lower in the final four time points.

Figure 5 displays the estimated difference curves comparing the optional and forbidden NT words under each preceding tone condition. The analysis reveals that forbidden NT words generally exhibit lower F0 contours than optional NT words. However, this difference reaches statistical significance only at specific time points: the final four time points after T1, the first two and last one time points after T2, and the last time point after T3. No significant effect of NT type was observed under the preceding T4 condition.

Interestingly, we also observed significant effects of the underlying tone under different preceding tones. Due to the limit of space, we focus on the results for underlying Tone 2 here: NT syllables with underlying T2 tend to have lower F0 contours

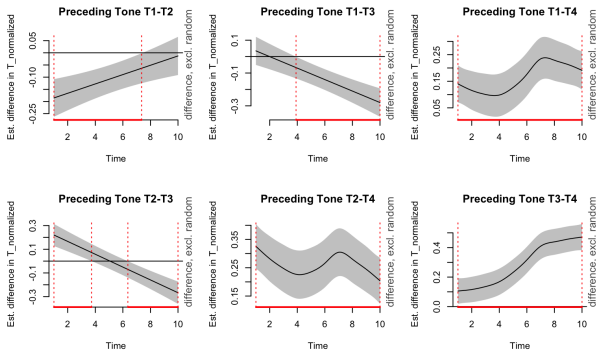


Figure 4: The estimated F0 difference trajectories between each two preceding tones.

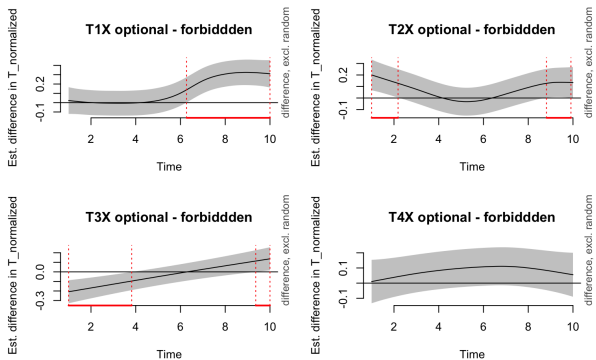


Figure 5: The estimated F0 difference trajectories between neutral tone types under different preceding tone conditions.

across preceding T1, T2, and T3, particularly in comparison to those with underlying T1 (see Figure 6). However, this effect disappears again following preceding T4. We found systematic effects for other underlying tones as well.

5. Discussion

This study investigates the acoustic variation of NT in Beijing Mandarin, focusing on two understudied types: optional and forbidden NT words.

Firstly, our study shows that the duration of NT syllables is approximately 77% of their preceding syllables, which is longer than the NT syllables reported in previous studies: 45% of their non-neutral counterparts according to [4, 1, 2], and 60% according to [5, 6]. Additionally, we find that forbidden NT words have a significantly longer duration than optional NT words. This suggests that both optional and forbidden NT words exhibit less reduction compared to obligatory NT words, with forbidden NT words being even less reduced than optional ones. This finding aligns with [7]’s argument that NT syllables vary in their degree of reduction rather than adhering to a strict binary distinction between “neutral” and “full” tones. A significant effect of NT type was also observed in F0 contours, further supporting the idea that different NT types exhibit distinct acoustic features.

Secondly, our results confirm that preceding tone plays a crucial role in shaping NT realization. We found that NTs following T1, T2, and T4 exhibit a mid-falling contour, with those

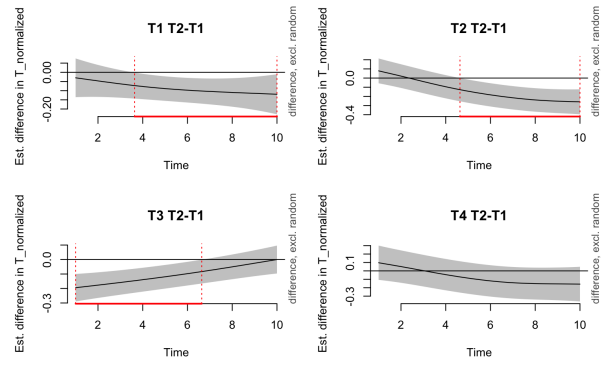


Figure 6: The estimated F0 difference trajectories between underlying Tone 2 and Tone 1 under different preceding tone conditions.

after T2 being highest, followed by those after T1, and the lowest after T4. NTs following T3 display a mid-level contour. These findings in general align with previous research [4, 5, 1, 2, 7]. The minor differences in the tone values may further suggest the effect of NT types.

However, this study also found significant effects of the underlying tone, contradicting previous research [12, 9, 10, 5, 6, 11]. Across different preceding tone conditions, we observed that the underlying tone systematically influenced NT realizations. These findings suggest that while NT contours are primarily conditioned by preceding tones, NT is not entirely toneless or without targets. Instead, it may exhibit different targets depending on its underlying tone, with these targets further shaped by other linguistic factors such as NT type. Further research is needed to determine whether this pattern is specific to more lexical NT words such as optional and forbidden NT.

Another interesting finding in our study is that both the effect of the underlying tone of NT syllables and the effect of NT type disappeared under the preceding T4 condition. This may suggest that T4 exerts the strongest conditioning effect on NT, overriding all other influences.

6. Conclusion

By analyzing the phonetic realization of NT—examining both duration and F0 contours—under the influence of three factors (preceding tone, underlying tone, and NT type), our findings confirm that preceding tones exert a dominant effect on NT. However, our results also reveal that NT is shaped by multiple interacting factors rather than merely functioning as a weak form. NT shows systematic variation across different underlying tone conditions, suggesting that phonetic targets may be influenced by the underlying tone. Furthermore, the distinction between optional and forbidden NT words supports the notion that NT exists on a continuum, with different types retaining varying degrees of phonetic properties. Future research should further investigate the effects of underlying tone across various NT syllable types and identify the acoustic features that best differentiate them. A deeper understanding of these variations is essential for refining theoretical models of NT and advancing our knowledge of prosodic variation in Mandarin and other languages.

7. References

- [1] W.-S. Lee, “A phonetic study of the neutral tone in beijing mandarin,” in *Proceedings of the 15th International Congress of Phonetic Sciences (ICPHS 2003)*. Barcelona., 2003, pp. 1121–1124.
- [2] W.-S. Lee and E. Zee, “Standard chinese (beijing),” *Journal of the International Phonetic Association*, vol. 33, no. 1, pp. 109–112, 2003.
- [3] J. Yang, Y. Zhang, A. Li, and L. Xu, “On the duration of mandarin tones,” in *Interspeech*, 2017, pp. 1407–1411.
- [4] M. Lin and J. Yan, “Beijingshua qingsheng de shengxue xingzhi,” *Dialect*, vol. 3, pp. 166–178, 1980.
- [5] J. Cao, “Putonghua qingsheng yinjie texing fenxi,” *Applied Acoustics*, vol. 5, no. 4, pp. 1–6, 1986.
- [6] Y. Chen and Y. Xu, “Production of weak elements in speech—evidence from f patterns of neutral tone in standard chinese,” *Phonetica*, vol. 63, no. 1, pp. 47–75, 2006.
- [7] J. Huang and A. Li, “Pu tong hua qing sheng fen xi yan jiu,” *Chinese Teaching in the World*, no. 3, pp. 369–387, 2023.
- [8] M. Hu, *Beijingshua chutan [A preliminary study of the Peking dialect]*. Beijing: Commercial Press, 1987.
- [9] Y. R. Chao, *Mandarin primer: An intensive course in spoken Chinese*. Harvard University Press, 1948.
- [10] —, *A grammar of spoken Chinese*. Berkeley: University of California Press, 1965.
- [11] J. Huang and F. Shi, “Han yu qing sheng yin jie yun lv biao xian de duo yang xing,” *Applied Linguistics*, no. 1, pp. 76–85, 2019.
- [12] M. J. Yip, “The tonal phonology of chinese,” Ph.D. dissertation, Massachusetts Institute of Technology, 1980.
- [13] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, “Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi,” in *Proc. Interspeech 2017*, 2017, pp. 498–502.
- [14] Y. Jadoul, B. Thompson, and B. de Boer, “Introducing Parselmouth: A Python interface to Praat,” *Journal of Phonetics*, vol. 71, pp. 1–15, 2018.
- [15] Y. Wang, A. Jongman, and J. A. Sereno, “Acoustic and perceptual evaluation of mandarin tone productions before and after perceptual training,” *The Journal of the Acoustical Society of America*, vol. 113, no. 2, pp. 1033–1043, 2003.
- [16] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2024, version 4.4.2. [Online]. Available: <https://www.R-project.org/>
- [17] E. R. Girden, *ANOVA: Repeated measures*. Sage, 1992, no. 84.
- [18] B. B. Douglas Bates, Martin Mächler and S. Walker, *Fitting Linear Mixed-Effects Models Using lme4*, 2015, vol. 67, no. 1. [Online]. Available: <https://doi.org/10.18637/jss.v067.i01>
- [19] S. N. Wood, “Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models,” *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 73, no. 1, pp. 3–36, 2011.
- [20] —, *Generalized Additive Models: An Introduction with R*, 2nd ed. CRC Press, 2017.