JUNE 01 2022

Asymmetrical roles of segment and pitch accent in Japanese spoken word recognition

Hironori Katsuda ; Jeremy Steffman

JASA Express Lett. 2, 065201 (2022)
https://doi.org/10.1121/10.0011573

JASA EXPRESS LETTERS ARTICLE asa.scitation.org/journal/jel

Asymmetrical roles of segment and pitch accent in Japanese spoken word recognition

Hironori Katsuda[1,a] and Jeremy Steffman[2]
[1]Department of Linguistics, UCLA, Los Angeles, California 90095, USA
[2]Department of Linguistics, Northwestern University, Evanston, Illinois 60208, USA
katsuda1123@gmail.com, jeremy.steffman@northwestern.edu

Abstract: This study examines the roles of segment and pitch accent in Japanese spoken word recognition. In a lexical decision task, it replicates the finding of Cutler and Otake [(1999) J. Acoust. Soc. Am. **105**(3), 1877–1888] that pitch accent restricts word activation with a more comprehensive, rigorous experimental design. Furthermore, results uncover an asymmetrical role of segment and pitch accent in word recognition in Japanese: words primed by a pitch accent-matching prime are recognized more slowly and less accurately than words primed by a segment-matching prime.

[Editor: Rachel M. Theodore] https://doi.org/10.1121/10.0011573

Received: 10 February 2022 **Accepted:** 13 May 2022 **Published Online:** 1 June 2022

1. Introduction

Both segmental contrasts and lexical prosody (i.e., word-stress, tone, pitch accent) play a crucial role in making lexical distinctions in language. A famous Mandarin Chinese example shows that the syllable *ma* bears four different meanings when combined with four different tones: *ma* with tone 1 (*ma1*) means "mother," while *ma2*, *ma3*, and *ma4* mean "hemp," "horse," and "to scorn," respectively. In Tokyo Japanese (henceforth "Japanese"), a pitch accent language, pitch accent patterns distinguish meanings as well. For example, the disyllable *hashi* means "chopsticks" with initial accent (*háshi*), "bridge" with final accent (*hashí*), and "edge" with no accent (*hashi*). Phonetically, the accent is realized as a steep fall in fundamental frequency beginning near the end of an accented mora (Beckman and Pierrehumbert, 1986).

The relative importance of suprasegmental and segmental information in word recognition has been examined in various studies with interesting results. These studies typically employ a lexical decision task and priming paradigm. Listeners are exposed to a prime, which is followed by a target item. Their task is to judge if the target is a word or not. Response time in this task is assumed to reflect the dynamics of lexical activation and selection, which will be impacted by the prime. For example, priming a word with itself (an "identity" prime) results in faster recognition (as compared to an unrelated prime), following from the activation of the word via priming, a facilitatory effect. Given this, consider a prime that matches in segmental information but *mismatches* in a suprasegmental feature (e.g., tone). To the extent that suprasegmental information matters in word recognition, this prime might vary in its facilitatory effects as compared to an identity prime. Following this logic and method, it has been shown that the role of suprasegmental information along these lines is language specific and heavily depends on its informativity for identifying words in the language. For example, a classic study by Cutler (1986) shows that lexical stress in English is not used to reduce the number of candidates in the early phase of word recognition (i.e., word activation). Specifically, members of a stress minimal pair that contrast solely in the location of stress, such as FORbear and forBEAR, are treated as homophones in the early phase, and the most likely candidate is picked out in the later phase. In other words, FORbear and forBEAR are equally facilitatory in priming, e.g., FORbear. This may be because, in English, the presence or absence of stress usually correlates with vowel quality (e.g., OBject ['ɑbd͡ʒɛkt] vs obJECT [əb'd͡ʒɛkt]), and thus stress information may not be critically used for word recognition. On the other hand, in Spanish and Dutch in which stress is not correlated to vowel quality, it has been shown that stress information *is* used to restrict word activation; a member of a stress minimal pair does not activate the other member via priming. Specifically, a word fragment prinCI- (from prinCIpio) does not activate (prime) PRINcipe in Spanish (Soto-Faraco *et al.*, 2001), and VOORnaam does not activate voorNAAM in Dutch (Cutler and Van Donselaar, 2001). However, Spanish and Dutch differ in how much stress information contributes to restrict word activation. In Spanish, stress is functionally phonemic, and stress mismatch is comparable to segmental mismatch, suggesting that stress contributes as much as segments do in word recognition. In Dutch, on the other hand, stress mismatch is not as detrimental as segmental mismatch, suggesting that stress contributes less than segments do.

[a]Author to whom correspondence should be addressed.

The role of tone in word recognition is commonly assessed in East-Asian tone languages, such as Mandarin Chinese (e.g., Lee, 2007; Poss *et al.*, 2008; Sereno and Lee, 2015), and it is shown that tonal information is used to restrict word activation. Sereno and Lee (2015) further show that while *segmental* primes (e.g., *ru3* → *ru4*) facilitate lexical decisions, tonal primes (e.g., *sha4* → *ru4*) slow down lexical decisions, suggesting that segmental mismatch is more detrimental than tonal mismatch to lexical access. Poss *et al.* (2008) focus on the inhibition effect of tone priming and argue that it is due to group activation of lexical items with the same lexical tone and competition among them during lexical selection.

The role of pitch accent has been studied in Japanese word recognition but is less studied compared to that of tone. Cutler and Otake (1999) show that pitch accent is used to restrict word activation, as the members of an accent minimal pair (e.g., *háshi* "bridge" and *hashí* "edge") do not prime each other (i.e., the lexical decision of the target primed by itself was significantly faster than that of the target primed by the other member of the accent minimal pair). However, the experimental design of Cutler and Otake's study is not completely ideal given the current understanding of this research area for the following three reasons. First, it does not distinguish finally accented and unaccented words; both are treated as the low-high pitch pattern (LH) words (e.g., *ame* "candy," *ichí* "one") as opposed to the high-low pitch pattern (HL) words (e.g., *áme* "rain," *íchi* "market"). Although finally accented and unaccented words are prosodically neutralized in isolation (Poser, 1984; Sugiyama, 2006), it is ideal to separate them and to be able to contrast all three accent patterns (i.e., initially accented, finally accented, and unaccented) to obtain a more comprehensive understanding of the role of pitch accent. Second, Cutler and Otake did not match word frequencies of the members of each pair. This might be problematic since Sekiguchi (2006) shows that word familiarity, which is highly correlated with word frequency, modulates the role of pitch accent in word activation. Finally, since their experimental design lacks a condition in which the members of each pair are matched in pitch accent but mismatched in segmental content (e.g., *ríka* "science" → *káme* "turtle"), the relative contribution of segmental and pitch accent cues in spoken word recognition is largely unknown. Accordingly, this study attempts to replicate and extend Cutler and Otake's study in a more comprehensive experimental design, which allows us to directly compare the roles of segment and pitch accent in Japanese word recognition.

2. Methods

To assess the relative importance of pitch accent and segmental information in word recognition, we implemented an auditory priming lexical decision task in which participants identified word and non-word targets as being lexical items or not. The target was primed by one of four prime types: a prime that matched the target in both segment and pitch accent (an *identity* prime), a prime that mismatched the target in both segment and pitch accent (a *control* prime), a prime that matched only in segmental material and mismatched in pitch accent (a *segment* prime), and a prime that matched in pitch accent but mismatched in segmental material (a *pitch accent* prime). Examples are shown in Table 1.

2.1 Predictions

If pitch accent restricts word activation, as Cutler and Otake (1999) show, we expect lexical decisions to be faster in the identity prime condition than in the segment prime condition. Furthermore, if the roles of pitch accent and segment are equivalent, as in the case of Spanish, we would observe no difference between the segment prime and pitch accent prime conditions. If the roles are asymmetrical, as in the case of Dutch and Mandarin Chinese, we would observe a difference between the two conditions, with a potential inhibitory effect of pitch accent-matching primes that is analogous to the inhibitory effect observed in tone-matching primes in Mandarin Chinese (Poss *et al.*, 2008; Sereno and Lee, 2015). We additionally assess listeners' accuracy, which we predict should be enhanced by activation in identity priming and potentially negatively impacted by inhibitory influences of pitch accent-matching primes.

2.2 Materials

We prepared 48 disyllabic word targets and the same number of disyllabic non-word targets. For each of the word targets (e.g., *káme* "turtle"), we prepared three types of semantically unrelated primes: a segment prime (*kamé* "pot"), a pitch accent prime (*ríka* "science"), and a control (*erí* "collar"), resulting in four prime conditions including the identity prime (*káme* "turtle"). The average log frequencies of the word targets, segment primes, pitch accent primes, and controls, based on word count in the Corpus of Spontaneous Japanese (Maekawa, 2003; Maekawa *et al.*, 2000) were 1.68, 1.63, 1.65, and 1.66, respectively. There was no significant difference among the four types of words [$F(3, 188) = 0.081$, $p = 0.97$]. The word targets are evenly divided into three accent patterns: 18 initially accented, 18 finally accented, and 18 unaccented. Both the segment primes and the controls are evenly distributed in terms of pitch accent pattern. For example, for 18

Table 1. Words exemplifying the different prime types used in the experiment.

Target	Identity	Segment	Pitch accent	Control
káme "turtle"	*káme* "turtle"	*kamé* "pot"	*ríka* "science"	*erí* "collar"

initially accented word targets, six segment primes were finally accented, while the other six were unaccented. To prosodically distinguish finally accented and unaccented words, words were embedded into a frame sentence *x-da* "(It) is *x*."

Since each target must be presented to each participant once, we created four counterbalanced lists by prime. Participants were randomly assigned to each of the four lists.

Stimuli were recorded by a native Tokyo Japanese speaker in a sound-attenuated booth, using an SM10A Shure™ (Niles, IL) microphone and headset. The stimuli were then manipulated using Praat (Boersma and Weenink, 2019). Specifically, the frame was spliced after each word to ensure invariant frame; however, since the pitch pattern of the frame varies depending on the pitch accent pattern of the word, this was done within each accent type, with the contextually appropriate rendition of *-da* spliced following each pitch accent pattern. The average fundamental frequencies of the syllable *da* were 91, 98, and 134 Hz for initially accented, finally accented, and unaccented words, respectively.

2.3 Participants and procedure

Prior to the onset of the COVID pandemic, we recruited 20 native speakers of the Tokyo dialect of Japanese. We additionally recruited 39 participants for remote participation, due to the necessity of pandemic-induced remote data collection. We recruited these remote participants using the platform Prolific. We subsequently excluded four participants who showed the poorest accuracy in their responses in the lexical decision task, each of whom showed under 60% accuracy of their responses in word trials, suggesting a misunderstanding of the task. Of these four excluded participants, one performed the task in person, and three performed the task remotely. In total, we analyzed responses from 55 participants (19 in person, 36 remote).

During the experiment, in-person participants were presented with audio stimuli binaurally via a Peltor™3M™ (Maplewood, MN) listen-only headset while seated in a sound-attenuated booth. These participants completed the experiment using a Millikey SH-2 Button Box (LabHackers Research Equipment, Halifax, Canada) linked to a laptop computer.

Remote participants were instructed to use their own headphones and take the experiment in a quiet room. Instead of a button box, remote participants provided responses via key press, using the "F" and "J" keys on their computer.

In each trial, participants hear a prime and target and make a lexical decision regarding the target. They were instructed to do so as quickly and accurately as possible by pressing the red (Yes) or green (No) buttons for in-person participants and the "J" (Yes) or "F" (No) keys for remote participants. The inter-stimulus interval between the prime and target was 250 ms. Prior to the test trials, participants completed eight practice trials. The experiment took approximately 15 min to complete.

2.4 Measures and analysis

We present two analyses here. First, we analyze reaction time for correct responses to word target trials (i.e., when a participant correctly identified a word as such, reflecting successful lexical access). Second, we present an accuracy analysis, assessing listeners' accuracy in identifying word targets. We excluded reaction time outliers from both analyses, dropping by-participant *z*-scored reaction times greater than 2.5 or smaller than –2.5. We analyzed the data using linear mixed effect Bayesian regression models, as implemented in *brms* (Bürkner, 2017). The dependent variable in the reaction time analysis was reaction time (in ms). We modeled the reaction times using an ex-Gaussian link function (exponentially modified Gaussian), specified as *family = exgaussian* in *brms*. The dependent variable in the accuracy analysis was a binomial correct/incorrect response; as such, we modeled accuracy responses with a logistic link function (as a binomial logistic regression model).

In both models, we examined how reaction time and accuracy varied by prime and by setting (in person vs remote), also including the interaction of these two fixed effects. We additionally included random intercepts for participant and by-participant slopes for prime. In R code, this is specified as RT/response ~ prime * setting + (1+prime|participant). Models were fit to draw 4000 samples in each of four Markov chains, discarding the first 1000 samples from each chain and keeping the remaining 75% of samples for inference. In assessing the effect of prime on reaction time, we report the following: the estimate and 95% credible interval (CrI) for an effect. In the Bayesian framework we employ, 95% CrI, which *excludes* zero, provides "credible" evidence for an effect, i.e., the model consistently estimates the same directionality, with a distribution that is unlikely to include zero. We also report the percentage of the posterior distribution for an estimate that shows a given directionality, using the p_direction function in *bayestestR* (Makowski *et al.*, 2019). This value ranges between 50 and 100 and is more intuitively compared to a frequentist *p*-value. For example, if 99% of the posterior shows a given sign, we can be confident in the directionality of the effect and would take this as reliable evidence for effect existence. We would indicate this as probability of direction (pd) = 99%. We consider *pd* values larger than 95% to indicate reliable evidence for an effect (note that when $pd > 97.5\%$, this corresponds to an exclusion of the value of zero in the 95% CrI in the posterior distribution). We additionally report marginal pairwise comparisons of interest for prime, which were extracted using *emmeans* (Lenth, 2020).

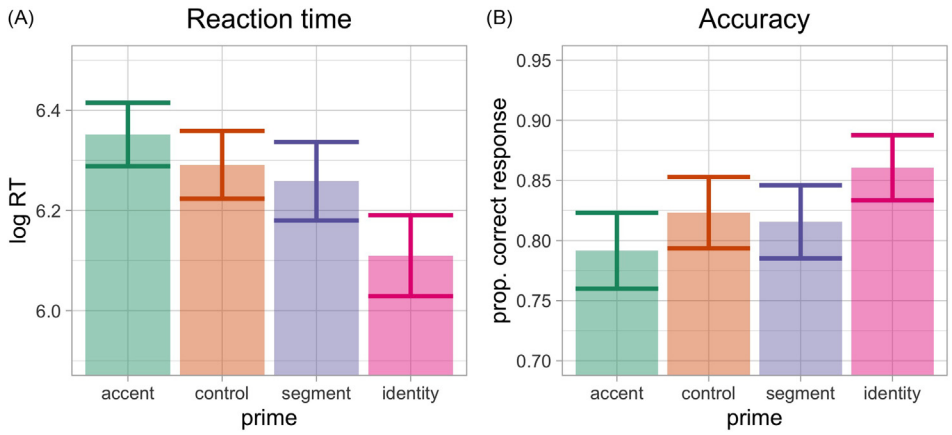

Fig. 1. Mean log-transformed reaction times (A) and accuracy (B) split by prime type. Error bars show 95% confidence interval (CI) computed from the raw data.

3. Results

3.1 Reaction time analysis

In coding the model contrasts in the reaction time model, we contrast-coded setting (in-person mapped to −0.5, remote mapped to 0.5). In coding the four-level variable prime, we mapped *segment primes* to the reference level, allowing us to examine estimates more easily for particular comparisons of interest. We specified weakly informative priors, with the intercept specified as normal (710, 50) (where 710 ms was the approximate mean log reaction time for the reference level in the model) and with fixed effect priors specified as normal (0, 100).

Figure 1(A) shows reaction time results, split by prime, while Table 2 shows the full estimates from the model. We report only the *pd* in the text. In assessing the results from the reaction time analysis, we consider how pairs of prime types relate to each other. We find that segment primes show reliably increased RT in relation to identity primes, i.e., matching in only segmental material (and *mismatching* in pitch accent) is detrimental as compared to matching in both ($pd = 100\%$). Segment primes are also not distinct from control primes in this regard ($pd < 95\%$). Furthermore, pitch accent primes show increased RT in relation to segment primes ($pd = 98\%$), suggesting that pitch accent priming is detrimental for RT beyond the slowdown evidenced from identity to segment primes. We also note that pairwise comparisons extracted with *emmeans* showed an expected difference between pitch accent and identity primes ($\beta = 70$, 95% CrI $= [40, 101]$, $pd = 100\%$) and between control and identity primes ($\beta = 63$, 95% CrI $= [34, 92]$, $pd = 100\%$), with faster reaction times to identity primes in both cases. The variable "setting" did not show a main effect ($pd < 95\%$), though we note that the estimates tend to be positive for the effect, suggesting overall participants in the remote setting had slower RTs. There was one credible interaction between identity primes (compared to the reference level) and setting. The negative estimate indicates faster RT for identity primes vs segment primes when the setting is remote as compared to in person, in other words, stronger identity priming for remote participants. However, for both settings, the effect shows the same directionality (faster responses to identity primes). Notably too, pairwise comparisons with *emmeans* of setting within each prime type (i.e., in person-segment primes to remote segment primes) found no reliable difference in RT across settings for any prime type, with all $pd < 95\%$. This, in addition to the lack of a main effect of setting, is taken to suggest that participants overall responded fairly comparably to primes across settings.

Table 2. Summary of the reaction time model. L-CrI and U-CrI refer to lower and upper 95% CrI. Fixed effects for which the 95% CrI excludes zero are in boldface.

	β (error)	L-CrI	U-CrI
(Intercept)	693 (25)	644	781
Prime, accent	**29 (14)**	**1**	**56**
Prime, control	23 (14)	−6	49
Prime, identity	**−41 (15)**	**−71**	**−11**
Setting	38 (36)	−33	109
Set: prime, accent	−13 (24)	−58	34
Set: prime, control	−26 (23)	−58	34
Set: prime, identity	**−63 (25)**	**−111**	**−12**

Table 3. Summary of the accuracy model. L-CrI and U-CrI refer to lower and upper 95% CrI.

	β (err)	L-CrI	U-CrI
(Intercept)	1.73 (0.16)	1.42	2.05
Prime, accent	−0.26 (0.17)	−0.60	0.06
Prime, control	−0.02 (0.17)	−0.35	0.31
Prime, identity	0.28 (0.19)	−0.08	0.67
Setting	−0.43 (29)	−0.99	0.15
Set: prime, accent	0.32 (0.30)	−0.28	0.93
Set: prime, control	0.26 (0.30)	−0.34	0.85
Set: prime, identity	0.59 (0.34)	−0.08	1.26

3.2 Accuracy analysis

The setting variable was coded in the same way as in the accuracy model. As before, we mapped segment prime to the reference level. We again used weakly informative priors specified as normal (1.49, 1) for the intercept (where 1.49 was the log-odds of a correct response at the reference level for prime) and normal (0, 1) for the fixed effects in log-odds space. Figure 1(B) shows the accuracy results, split by prime, while Table 3 shows the estimates from the model.

The accuracy analysis finds that, with segment primes as the reference level, no prime effects are credible (all $pd < 95\%$), though we can note essentially no difference is estimated between segment and control primes ($\beta = 0.02$), while identity primes show a positive (more accurate) estimate ($\beta = 0.28$), and accent primes show a negative (less accurate) estimate ($\beta = -0.26$). Using *emmeans* again to compare across all prime types, only one pairwise difference is credible: the difference between identity and accent primes, with accent primes showing credibly decreased accuracy in comparison to identity primes ($\beta = -0.53$, 95% CrI = [−0.94, −0.18], $pd = 100\%$). In other words, accent-primed words are recognized less accurately compared to identity-primed words, while there is otherwise no credible difference in accuracy across prime types.

4. Conclusions

In summary, we find that words primed by a segment-matching prime are recognized more slowly than an identity prime, replicating Cutler and Otake's finding in a more comprehensive experiment. This reinforces the claim that pitch accent information is indeed important in restricting lexical activation when segmental material supports multiple lexical hypotheses. We further find that words primed by a pitch accent-matching prime are recognized more slowly than words primed by a segment-matching prime. Pitch accent priming also reduces accuracy, whereas accuracy is not negatively impacted by segment priming (as compared to identity priming). As such, we have novel evidence for an asymmetrical role of segmental material and suprasegmental material (pitch accent) in word recognition in Japanese.

We thus conclude that while pitch accent information is useful in restricting activation when segmental information alone is insufficient (as shown by RT differences between segment and identify primes), pitch accent-matching alone, in the absence of segmental overlap, is detrimental. This inhibitory effect is analogous to the effect observed in the tone-matching prime condition in Mandarin Chinese (Poss *et al.*, 2008; Sereno and Lee, 2015). Following Poss *et al.* (2008), we attribute this to the group activation hypothesis: pitch accent priming evokes lexical items with the same pitch accent status, and competition among them results in the inhibitory effect observed in the pitch accent-matching condition. In this sense, we have support for the claim that segmental information in Japanese plays a dominant role in lexical processing, as it allows the listener to narrow their consideration of lexical hypotheses more efficiently as compared to pitch accent alone which, in the lens of group activation, activates a large cohort of pitch accent-matching candidates. These results thus offer a new lens into the role of suprasegmental information in spoken word recognition in Japanese. Future work will benefit from considering how these results generalize to more complex prosodic/intonational contexts that may contextually determine pitch accent realization and to other languages that use suprasegmental cues to lexical identity in similar fashion to Japanese.

Acknowledgments

Many thanks to Sun-Ah Jun, Megha Sundara, Pat Keating, Claire Moore-Cantwell, and members of the UCLA Phonetics Laboratory for valuable feedback on this study.

References and links

Beckman, M., and Pierrehumbert, J. (**1986**). "Intonational structure in Japanese and English," Phonol. Yearb. **3**, 255–309.

Boersma, P., and Weenink, D. (**2019**). "Praat: Doing phonetics by computer (version 6.1.05) [computer program]," http://www.praat.org/ (Last viewed 5/25/2022).

Bürkner, P. (**2017**). "brms: An R package for Bayesian multilevel models using Stan," J. Stat. Softw. **80**(1), 1–28.
Cutler, A. (**1986**). "Forbear is a homophone: Lexical prosody does not constrain lexical access," Lang. Speech **29**(3), 201–220.
Cutler, A., and Otake, T. (**1999**). "Pitch accent in spoken-word recognition in Japanese," J. Acoust. Soc. Am. **105**(3), 1877–1888.
Cutler, A., and Van Donselaar, W. (**2001**). "Voornaam is not (really) a homophone: Lexical prosody and lexical access in Dutch," Lang. Speech **44**(2), 171–195.
Lee, C. Y. (**2007**). "Does horse activate mother? Processing lexical tone in form priming," Lang. Speech **50**(1), 101–123.
Lenth, R. (**2020**). "emmeans: Estimated marginal means, aka least-squares means. R package version 1.5.3," https://CRAN.R-project.org/package=emmeans (Last viewed 5/25/2022).
Maekawa, K. (**2003**). "Corpus of spontaneous Japanese: Its design and evaluation," in *Proceedings of the ISCA and IEEE Workshop on Spontaneous Speech Processing and Recognition*, April 13–16, Tokyo, Japan, pp. 7–12.
Maekawa, K., Koiso, H., Furui, S., and Isahara, H. (**2000**). "Spontaneous speech corpus of Japanese," in *Proceedings of the Second International Conference of Language Resources and Evaluation*, May 31–June 2, Athens, Greece, pp. 947–952.
Makowski, D., Ben-Shachar, M., and Lüdecke, D. (**2019**). "bayestestR: Describing effects and their uncertainty, existence and significance within the Bayesian framework," J. Open Source Softw. **4**(40), 1541.
Poser, W. (**1984**). "The phonetics and phonology of tone and intonation in Japanese," Ph.D. dissertation, MIT, Cambridge, MA.
Poss, N., Hung, T. H., and Will, U. (**2008**). "The effects of tonal information on lexical activation in Mandarin," in *Proceedings of the 20th North American Conference on Chinese Linguistics (NACCL-20)*, April 25–27, Columbus, OH, Vol. 1, pp. 205–211.
Sekiguchi, T. (**2006**). "Effects of lexical prosody and word familiarity on lexical access of spoken Japanese words," J. Psycholinguist. Res. **35**(4), 369–384.
Sereno, J. A., and Lee, H. (**2015**). "The contribution of segmental and tonal information in Mandarin spoken word processing," Lang. Speech **58**(2), 131–151.
Soto-Faraco, S., Sebastián-Gallés, N., and Cutler, A. (**2001**). "Segmental and suprasegmental mismatch in lexical access," J. Mem. Lang. **45**, 412–432.
Sugiyama, Y. (**2006**). "Japanese pitch accent: Examination of final-accented and unaccented minimal pairs," Toronto Work. Papers Linguist. **26**, 73–88.