Prosodic prominence in Greek: methodological and theoretical considerations

Riccardo Orrico, Stella Gryllia, Na Hu, Jiseung Kim, Amalia Arvaniti

Radboud University, The Netherlands

riccardo.orrico@ru.nl, stella.gryllia@ru.nl, na.hu@ru.nl, jiseung.kim@ru.nl, amalia.arvaniti@ru.nl

Abstract

A popular paradigm for studying prominence is Rapid Prosody Transcription (RPT), in which linguistically untrained participants listen to utterances and mark on a transcript words they perceive as prominent. RPT responses can be sensitive to the type of instructions used, such as whether participants are asked to select only the most prominent word or all the words they deem prominent. Here, we compare results from two RPT studies with the same Greek materials but using the two above-mentioned instruction types. Inter-rater agreement scores were similar across the two studies and yielded comparable overall results (e.g., words in focus were more likely to be selected in both). However, the relevance of certain criteria varied depending on task: in the multi-word task, duration, amplitude, and $F0_{max}$ predicted prominence, while in the single-word task $F0_{max}$ was not a predictor. These differences suggest that the tasks investigated here are not interchangeable, despite similarities. Most importantly, they can each lead to different interpretations of what constitutes prominence, suggesting that researchers need to be cautious when using them.

1. Introduction

A number of studies in recent years have investigated the properties that make some words stand out compared to others. These studies have shown that various factors contribute to prominence assessment: acoustic cues (e.g., F0, duration, and amplitude), phonological properties (e.g., presence and type of pitch accent), semantic-pragmatic information (e.g., information structure), lexical factors (e.g., part of speech, lexical frequency), and syntactic factors (e.g., canonical vs. cleft sentences); among others, see [1], [2], [3], [4], [5], and [6].

Despite numerous studies, there is still a lack of agreement as to which cues contribute to prominence, and what their relative contribution might be, even when the same or similar languages are tested. This ultimately affects our understanding of prominence, a point we return to in §4. For example, some studies show that changes in F0 are strongly related to prominence [2], [3], while others have reported a weaker relation between the two [1], [7]. Some of these inconclusive results may be linked to methods used to study prominence. In the last decade, many studies have relied on the Rapid Prosody Transcription paradigm (henceforth RPT, [1], [8]), in which linguistically untrained participants listen to utterances and mark on a transcript phrasal boundaries and prominent words. An increasing number of languages have been investigated using RPT, including prominence perception by non-native listeners [9] and listeners unfamiliar with the language of the stimuli [10]. Despite the popularity of RPT, some studies suggest that responses are sensitive to the instructions used and the language tested. For instance, [11] instructed one group of English-speaking participants to pay attention to acoustic cues, and another group to focus on meaning. The same instructions were tested with English, Spanish, and French listeners [4]. Both [4] and [11] report that the instructions steered listener preferences towards a specific set of cues, though the effect was smaller in English than French and Spanish [4].

Given the sensitivity of RPT to language and task, we conducted two RPT studies using different instructions to test their suitability for Greek. The two versions differed as to the number of words listeners were asked to select, one word per utterance (*single-word task*), as in [5], or as many as they saw fit (*multi-word task*), as in [1] and many other studies. These two versions were chosen for the following reasons. First, in Greek, all content words carry a pitch accent, unless they occupy a post-nuclear position [12]. This could potentially make it difficult for participants to differentiate degrees of prominence. An additional concern was that in Greek, the same term *τόνος* [ˈtonos] is used to denote both orthographic accent and the phenomena of *stress*, *accent*, and *prominence*. Further, all polysyllabic words are orthographically accented, but the orthographic accent does not always correspond to what we could call prominence; e.g., monosyllabic verbs (e.g., ζω [zo] “I live”, which most likely carry a pitch accent, do not bear orthographic accent, but disyllabic function words (e.g. από [apo] “from”), which are unaccented in running speech, do [12]. Thus, if Greek participants were asked to select [toniˈzmenes ˈleksis], which translates to both “prominent words” and “orthographically accented words”, they could opt for all the words with orthographic accent. If so, instructions to select only the most prominent word could be advantageous, as they would help clarify our intended meaning. Additionally, a comparison between the two instruction types allowed us to investigate the extent to which they lead to different conclusions regarding prominence assessment. To address this point, we tested the two versions with comparable groups of Greek participants. Their responses were analyzed as a function of a predefined set of predictors shown to affect RPT responses across languages, i.e., phonological properties related to the presence and type of pitch accent, and acoustic properties of words, to test for the effect of instructions. Thus, our aim was not to conduct a traditional RPT study, in which the responses are analyzed to observe what makes words prominent, but to test the suitability of the two tasks in Greek. This, in turn, allowed us to make observations about RPT and the notion of prominence that go beyond Greek.

2. Method

2.1. Participants

Twenty native speakers of Greek participated in the *multi-word* study: 13 F, 5 M, 1 non-binary, 1 preferred not to say; 20–61 age range, $\bar{x}$ = 31.50, SD = 13.1. Another 17 took part in the *single-word* study: 14 F, 2 M, 1 non-binary; 19–59 age range $\bar{x}$ = 30, SD = 12.81. None of the participants reported any history of speaking or hearing disorders. Participants were recruited via personal networks and participated voluntarily. Results from power analysis conducted on a separate RPT study in Greek with a similar design suggest that the number of participants is adequate for the type of analysis we report [13].

2.2. Description and analysis of the stimuli

The same 74 utterances, recorded by a 46 y.o. female native speaker of Greek were used as stimuli in both studies. They ranged from 2 to 8 orthographic words (277 in total) and included different sentence types (see Table 1). Two study authors listened to the utterances to ensure they were produced with the intended tune, focus location, and phrasing (one intonational phrase consisting of one intermediate phrase) [12]. Broad focus statements (16 items) were realized with a H* nuclear accent on the final word and L*+H accents on all preceding content words. Narrow focus statements were realized with a L+H* accent on the focused word and prenuclear L*+Hs on all preceding content words (final narrow focus, N = 10; non-final narrow focus, N = 16). All statements ended in L-L%. Polar questions were realized with a L* accent on the focused word, L*+Hs on all preceding content words, and H-L% edge tones (final focus, N = 10; non-final focus, N = 16). Finally, wh-questions (6 items) had a nuclear L*+H accent on the sentence-initial wh-word and L-H% edge tones. For more details on Greek prosodic structure, see [12].

Table 1: *Examples and tunes of the stimuli; nuclear accents and words bearing them are in bold.*

Sentence Type	**Tune**	**N**
Broad focus statements	(L*+H) **H*** L-L%	16
[ˈeçi ˈkrio neˈro sto **psiˈɟjio**]		
“There’s cold water in the **fridge**.”		
Narrow focus statements	(L*+H) **L+H*** L-L%	26
[to **ɣaˈlazjo** ˈforema]		
“The **blue** dress.”		
Polar questions	(L*+H) **L*** H-L%	26
[sto ˈmano **aˈresi** to θimaˈrisço ˈmeli]		
“Does Manos **like** thyme honey?”		
Wh-questions	**L*+H** L-H%	6
[ˈ**pote** na poˈtiso ta luˈluðja]		
“**When** should I water the flowers?”		

We first analyzed the duration, Root Mean Square (RMS) amplitude, and F0$_{max}$ of individual words. For polysyllabic content words, the values were extracted from their stressed syllable; function words were always monosyllabic. The values, extracted in Praat [14], were z-scored and used as dependent variables in linear mixed models in R [15]. The models had the phonological status of the word (6 levels: unaccented, H*, L*, L+H*, L*+H, prenuclear L*+H) as fixed factor and the utterance as random intercept. We note that the results are not the output of a full production study but only serve to describe the stimuli; they should be treated with some caution due to the small sample elicited from one speaker.

Overall, these analyses show that the acoustic properties of the words varied substantially across phonological categories. First, accented words were not always more salient acoustically (*viz*. louder, longer, or higher pitched) than unaccented ones. Second, there were differences related to accent type. Accented words were longer than unaccented words, with the exception of those with L*+H (Fig. 1A). Additionally, words with L+H* had the highest RMS of all categories except for words carrying prenuclear accents (Fig. 1B). Finally, as expected, L* had significantly lower F0$_{max}$ than the other categories, while rising accents (L+H* and L*+H) had the highest F0$_{max}$ (Fig. 1C).

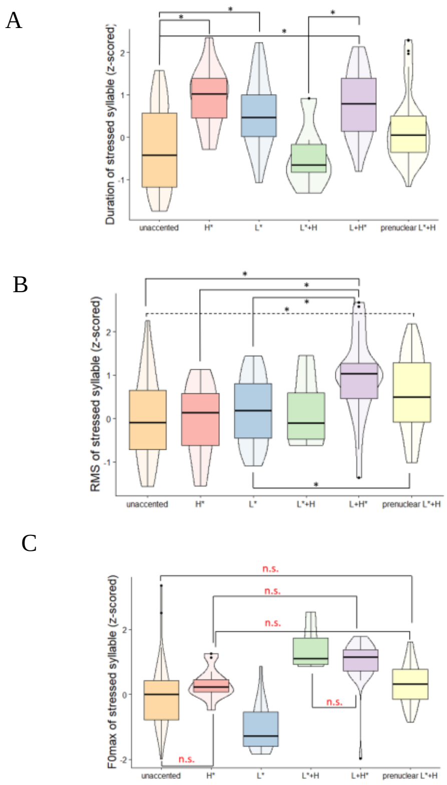

Figure 1: *z-scored Duration (A), RMS amplitude (B), and F0$_{max}$ (C) of stressed syllables in the stimuli.*

2.3. Procedure

The two studies were conducted online. Both consisted of 3 practice trials followed by 74 experimental trials, and included 4 prompts for self-paced breaks. Trial randomization was done by creating three lists containing all the stimuli in different orders; the lists were randomly assigned to participants. In each trial, participants listened to an utterance twice while seeing its Greek transcript on screen; the transcript lacked punctuation, other than apostrophes in contractions, and capitalization, except for proper names. During the second repetition, a checkbox appeared next to each orthographic word. In both studies, participants were asked to select the word(s) that in their opinion “stand out from the rest, that is, sound more important or stressed or as if the speaker has emphasized them” by checking the box next to that word. The wording of the instructions was chosen to overcome the terminology issue discussed in §1 and avoid biasing participants towards focusing on either acoustic or semantic aspects of the words [cf. 4, 11]. Participants in the multi-word study were instructed to select as many words as they saw fit, while participants in the single-word study were told to select only the most prominent one.

2.4. Statistical analysis

Following previous RPT studies ([2], [4]), we calculated *Fleiss’ kappa (κ)* and *p-scores*. Fleiss’ κ (which ranges from 0 to 1, with 1 representing perfect agreement) was calculated for all utterances as an overall indication of inter-rater agreement, and separately by sentence type. P-scores, percentages of participants who mark a word as prominent, were used for visualization purposes. We also ran Generalized Linear Mixed Effect Models (GLMMs) separately for each study. The dependent variable was the RPT binary response (1 selected; 0 not selected). The independent variables related to phonological and acoustic properties of the words (see §2.2): (i) metrical strength (unaccented, prenuclear accent, nuclear accent); (ii) nuclear accent type (H*, L*, L*+H, L+H*); (iii) duration; (iv) RMS amplitude; and (v) $F0_{max}$. Following [4], we ran five separate models, one for each independent variable. The random structure included intercepts for participants and for word nested within the utterance.

3. Results

3.1. Inter-rater agreement

No substantial differences in inter-rater agreement were detected between studies: Fleiss' κ was 0.40 (z = 90.1) for the multi-word task, and 0.38 (z = 58.6) for the single-word task. These scores indicate *fair* to *moderate* agreement as has also been reported for English RPT [1]. As shown in Table 2, however, Fleiss' κ varied considerably between tasks with respect to specific sentence types, such as narrow focus statements, while variation across sentence types was somewhat greater within the multi-word task.

Table 2: *Fleiss' κ scores with z in parentheses; all κ values are significantly different from zero.*

Sentence Type	κ (multi)	κ (single)
Broad Focus Statements	0.34 (36.9)	0.31 (23.3)
Narrow Focus Statements	0.46 (59.5)	0.39 (34.5)
Polar questions	0.43 (55.7)	0.42 (38.5)
Wh-questions	0.25 (17.6)	0.30 (14.5)

3.2. Phonological predictors

Fig. 2 shows that the two tasks yielded very similar p-scores with respect to the phonological predictors, viz. presence and type of accent. The GLMMs confirmed this impression.

With respect to metrical strength, the models showed that accented words were more likely to be selected as prominent than unaccented words, and words with nuclear accents were selected more often than those with prenuclear accent. This applied to both tasks and all pairwise comparisons: for the multi-word task, *unaccented vs. prenuclear*: 2.13 (.19), z = 11.23; *unaccented vs. nuclear*: 3.67 (.18), z = 20.04; *prenuclear vs. nuclear*: 1.54 (.21), z = 7.24; for the single-word task: *unaccented vs. prenuclear*: 2.05 (.27), z = 7.72; *unaccented vs. nuclear*: 3.25 (.25), z = 12.85; *prenuclear vs. nuclear*: 1.21 (.23), z = 5.19; p < .0001 in all cases.

The models testing the type of nuclear pitch accent revealed no significant differences among words accented with L*, L+H*, and L*+H. Only those carrying H* were significantly less likely to be selected as prominent relative to those with other accents, and this applied to both tasks.

3.3. Acoustic predictors

The models showed a positive correlation between stressed syllable duration and the likelihood of a word being selected as prominent. This applied to both studies, though the estimate value was larger for the multi-word task (0.95 (0.11), z = 8.15) than the single-word task (0.65 (0.14), z = 4.53; p < .0001 for both. Similarly, the higher its RMS amplitude, the more likely a word was to be selected, but with greater magnitude for the multi-word task (multi-word: 0.96 (0.12), z = 7.93; single-word: 0.66 (0.15), z = 4.46); p < .0001 for both. $F0_{max}$ was significant only in the multi-word task (0.38 (0.13), z = 3.02, p < .01; single-word task (0.10 (0.15), z = 0.65, p > .05).

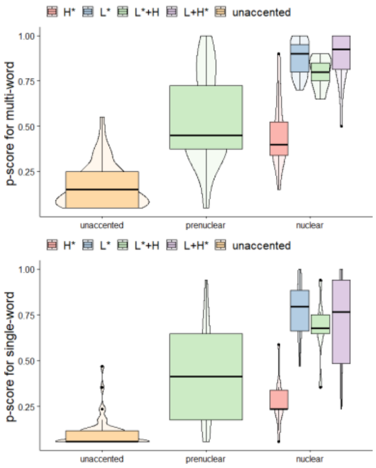

Figure 2. *Distribution of p-scores as a function of phonological predictors (presence and type of pitch accent) for the multi-word task (top) and single-word task (bottom).*

4. Discussion and conclusions

We conducted two RPT studies in Greek using two types of instruction, one asking participants to mark as prominent as many words as they saw fit (*multi-word task*), the other asking them to select only the word they considered the most prominent (*single-word task*). Overall, the two tasks did not yield substantially different outcomes. This is reflected in their Fleiss' κ scores, which indicate comparable overall inter-rater agreement. Essentially, the words selected by participants in the single-word task were to a large extent a sub-group of the words selected by participants in the multi-word task. With respect to cues used, in both studies, responses were largely affected by metrical strength, i.e., whether the word was accented or not, and whether the accent was nuclear or prenuclear. In contrast, accent type was a weaker predictor, as all nuclear accents, except H*, were equally likely to be selected as prominent. Overall, our results provide evidence that both tasks are viable in Greek and confirm the suitability of the single-word task for RPT, also shown in [5] with English and Samoan participants.

Further, our findings have repercussions for our understanding of prominence in Greek, the nature of RPT and of prominence more generally. We address each in turn.

The similarities between the two studies help us understand how Greek listeners interpreted prominence regardless of potential task differences. As noted, metrical structure was a strong predictor of prominence assessment. This result differs from findings of previous studies on English and German, which have shown that the relative prominence of different accents depends on their F0 properties: falling and low accents are typically rated less prominent than high accents, and high accents are rated less prominent than rising accents [4], [6], [16], [17]. Our results do not support these earlier findings: in Greek, L*s were as likely as L*+Hs and L+H*s to be selected as prominent when all were found in nuclear position. This suggests that our Greek listeners perceived as prominent the words in strong metrical positions regardless of their acoustic properties (see also [5] and [18]). The only exception was H*: H*-accented words were less likely to be selected as prominent relative to all other accents, both nuclear and prenuclear. A plausible reason that in Greek, L* and L+H* narrow the focus domain to one word, which therefore stands out metrically relative to others. In contrast, H* marks the last word in "broad focus" statements, as in other languages, such as English. Thus, our results for H* provide evidence in support of [19] who argue that broad focus sentences do not have focus. In short, words bearing H*s in Greek did not stand out relative to other words because they marked the last item in broad focus declaratives. This result points to the fact that listeners do not always consider as prominent words that are acoustically salient: as shown in §2.2, the stressed syllables of H*-accented words were not acoustically less salient than those of words bearing other accents (see Fig. 1). Rather, our participants decided based not on acoustic salience but higher order criteria, both phonological and pragmatic (cf. [5], [6], [15], [20]).

Further, the relative importance of the criteria used by our participants differed by task. As mentioned, the relation between responses and *phonological* predictors was similar across tasks. These similarities indicate that the connection between phonological cues and prominence is strong and unlikely to shift as a function of task. On the other hand, the relationship between prominence and *acoustic* predictors showed notable cross-task effects. This was most evident with respect to $F0_{max}$, which was a significant predictor only in the multi-word task. Perhaps this outcome is not surprising: since nuclear accents were those predominantly selected as prominent in the single-word task and they included L*, $F0_{max}$ was an unlikely predictor [cf. 18]. However, the same applies, to an extent, to duration and amplitude for which no such explanation is possible: though both duration and amplitude correlated positively with the probability of word selection, the magnitude of the effect was larger in the multi-word task. One interpretation of these differences is that in the multi-word task, our participants paid more attention to acoustic salience, while in the one-word task, they focused on metrical structure.

These findings have repercussions for how we interpret RPT results. As mentioned, while both the multi- and single-word task proved viable, our participants used the acoustic predictors more in the former than the latter. This suggests that caution is needed when raw acoustic values are treated as prominence transducers, because the conclusions we reach about them are task-related. Specifically, we argue that differences like the F0 effect in our two tasks are directly related to task demands: a higher number of predictors is likely to be statistically significant when participants are asked to select more words; this is because allowing them to select more words implicitly allows them to rely on more criteria. In turn, this runs the risk of one study claiming F0 to be a predictor of prominence in a given language, [2], [3], while another argues for the opposite, [1], [7].

At a minimum, the task differences we report suggest that while both tasks are useful for investigating prominence, they cannot be used interchangeably: one or the other should be selected depending on the specific research question addressed by the study. One could argue that the multi-word task yields a finer-grained picture, by allowing researchers to capture the effect of predictors that make modest contributions to prominence. The role of these predictors may be minimized or even eliminated in the single-word task. Thus, the multi-word task is more suitable for studies with a broad scope, such as the investigation of all potential cues to prominence in order to determine their relative importance at the group or individual level; [2], [6]. The single-word task may be useful for preliminary investigations of prominence and for obtaining a general understanding of the link between prominence perception and linguistic factors. It may also be a better choice for testing a specific hypothesis, such as whether L* is considered less prominent than H* by the speakers of a given language. The single-word task, however, may require greater control of the stimuli: it may not be suitable for long utterances or utterances with more than one intonational phrase, as such stimuli could be particularly challenging for lay participants when they have to choose only one prominent word.

At any rate, while RPT is a viable paradigm yielding replicable results (see, for example, [15] and [6]), multiple studies, including the present one, indicate that its output is sensitive to the types of instructions used and the language(s) under investigation. This has clear repercussions for the way we relate the results—e.g., the significance of predictors—to the notion of prominence. If we ascribe to a purely psychophysical view of prominence that can be investigated *post hoc* by asking participants to mark prominent words and then examining their characteristics, we may reach different conclusions about prominence in a given language, depending on the instructions used or, potentially, other aspects of the study's set up. This may not be much of an issue in the Germanic languages that have been mostly tested with RPT, as in these languages, the concept of phrasal stress is relatively clear to study participants, while acoustic salience and metrical strength largely go hand in hand, as argued in [18]. In other languages, however, this may not apply. This is one way to interpret the fact that instructions influence performance more in languages like French, in which Germanic-style prominence does not apply [4], or the greater variability evinced in the responses of Samoan speakers in [5]. Our Greek results, which show that accentuation trumps pitch height, add to this trend. Thus, caution should be applied when using RPT to determine what prominence is in a given language, and what we understand as prominence more generally. We note that the answer to this last question cannot simply be that prominence is expressed in different ways across languages, as this line of argumentation leads to an understanding of prominence that cannot generate constrained predictions and is essentially unfalsifiable. Instead, we side with [18] in asserting the preponderance of metrical strength, as evinced by our results.

To conclude, our studies showed that RPT is a robust paradigm suitable for investigating prominence in Greek despite our initial concern about the lack of distinct terms for *orthographic accent*, *accent*, and *stress*. Nevertheless, our findings also indicate that task instructions can affect responses, by drawing participants' attention to different dimensions of the signal. This suggests that researchers need to be cautious when selecting RPT and when interpreting the results, especially when investigating new languages.

5. Acknowledgments

This work was supported by the European Research Council through grant ERC-ADG-835263 (SPRINT) to A. Arvaniti. We thank our participants, our Erasmus interns, Vasiliki Kyritsi and Vasiliki Barouta, for their assistance with data collection, and Bob Ladd and Jennifer Cole for many helpful discussions on this work and prominence more generally.

6. References

[1] Cole, J., Mo, Y., & Hasegawa-Johnson, M. (2010). Signal-based and expectation-based factors in the perception of prosodic prominence. *Laboratory Phonology*, 1, 425–452. DOI: http://doi.org/10.1515/labphon.2010.022

[2] Baumann, S., & Winter, B. (2018). What makes a word prominent? Predicting untrained German listeners' perceptual judgments. *Journal of Phonetics*, 70, 20–38. DOI: https://doi.org/10.1016/j.wocn.2018.05.004

[3] Bishop, J., Kuo, G., & Kim, B. (2020). Phonology, phonetics, and signal-extrinsic factors in the perception of prosodic prominence: Evidence from Rapid Prosody Transcription. *Journal of Phonetics*, 82, 100977.

[4] Cole, J., Hualde, J. I., Smith, C. L., Eager, C., Mahrt, T., & de Souza, R. N. (2019). Sound, structure and meaning: The bases of prominence ratings in English, French and Spanish. *Journal of Phonetics*, 75, 113-147. DOI: https://doi.org/10.1016/j.wocn.2019.05.002

[5] Calhoun, S., Wollum, E., & Kruse Va'ai, E. (2021). Prosodic prominence and focus: Expectation affects interpretation in Samoan and English. *Language and Speech*, 64(2), 346-380. DOI: https://doi.org/10.1177/002383091989036

[6] Orrico, R., Gryllia, S., Kim, J.K. & Arvaniti, A. (2023). The influence of empathy and autistic-like traits in prominence perception. In R. Skarnitzl & J. Volín (Eds.), *Proceedings of the 20th International Congress of Phonetic Sciences* (pp. 1280-1284). GUARANT International.

[7] Kochanski, G., Grabe, E., Coleman, J., & Rosner, B. (2005). Loudness predicts prominence: Fundamental frequency lends little. *Journal of the Acoustical Society of America*, 118(2). 1038–1054

[8] Cole, J., & Shattuck-Hufnagel, S. (2016). New methods for prosodic transcription: Capturing variability as a source of information. *Laboratory Phonology*, 7(1): 8, 1–29. DOI: http://doi.org/10.5334/labphon.29

[9] Pintér, G., Mizuguchi, S., & Tateishi, K. (2014). Perception of prosodic prominence and boundaries by L1 and L2 speakers of English. In *Fifteenth Annual Conference of the International Speech Communication Association*.

[10] Riesberg, S., Kalbertodt, J., Baumann, S., & Himmelmann, N. P. (2020). Using Rapid Prosody Transcription to probe little-known prosodic systems: The case of Papuan Malay. *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, 11.

[11] Cole, J., Mahrt, T., & Hualde, J. I. (2014). Listening for sound, listening for meaning: Task effects on prosodic transcription. In *Proceedings of Speech Prosody* (Vol. 7, pp. 859-863).

[12] Arvaniti, Amalia & Baltazani, Mary. (2005). Intonational analysis and prosodic annotation of Greek spoken corpora. In Jun, S. A. (Ed.). *Prosodic Typology: The Phonology of Intonation and Phrasing*. 84-117.

[13] Orrico, R., Gryllia, S., Arvaniti, A. (*in preparation*) Prominence depends on phonology, not acoustic salience.

[14] Boersma, P., & Weenink, D. (2023). Praat: doing phonetics by computer [Computer program]. Version 6.3.13, retrieved 31 July 2023 from http://www.praat.org/

[15] R Core Team (2023). _R: A Language and Environment for Statistical Computing_. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>.

[16] Arvaniti, A., Gryllia, S., Zhang, C., Marcoux, K. P. 2022. Disentangling emphasis from pragmatic contrastivity in the English H*~L+ H* contrast. *Proceedings of Speech Prosody 2022*. doi: 10.21437/SpeechProsody.2022-170

[17] Baumann, S., & Röhr, C. (2015). The perceptual prominence of pitch accent types in German. In The Scottish Consortium for ICPhS 2015 (Ed.), *Proceedings of the 18th international congress of phonetic sciences* (paper number 0298.1-5). Glasgow, UK: The University of Glasgow.

[18] Ladd, D. R., & Arvaniti, A. (2023). Prosodic prominence across languages. *Annual Review of Linguistics*, 9, 171-193.

[19] Katz, J., & Selkirk, E. (2011). Contrastive focus vs. discourse-new: Evidence from phonetic prominence in English. *Language*, 771-816.

[20] Im, S., Cole, J. & Baumann, S., (2023) Standing out in context: Prominence in the production and perception of public speech, *Laboratory Phonology* 14(1). doi: https://doi.org/10.16995/labphon.641