EVIDENCE FOR THE COMPOSITIONALITY OF INTONATIONAL MEANING

Stella Gryllia[1,2], Mary Baltazani[3], Amalia Arvaniti[4]
[1]University of Leiden, [2]Utrecht University [3]University of Oxford [4]University of Kent
[1]s.gryllia@hum.leidenuniv.nl [2]mary.baltazani@ling-phil.ox.ac.uk [3]a.arvaniti@kent.ac.uk

ABSTRACT

The pragmatic interpretation of Greek wh-questions with different intonation was tested, by asking participants to listen to questions and bet on two follow-up sentences offering alternative explanations on the question's purpose (information- or non-information-seeking). L*+H L-!H% and L+H* L-L% were used and crossed (L+H* L-!H% and L*+H L-L%), giving rise to four experiment versions in a between-participant design. Responses from 190 Greek listeners supported previous analyses according to which L*+H L-!H% and L+H* L-L% lead to a preference for information-seeking vs. non-information seeking interpretations respectively. Responses were affected by both the pitch accent and boundary tone, with the joint contribution being most evident in the "crossed" tunes (L+H* L-!H% and L*+H L-L%). These results also support the notion that accents and edge tones contribute independently to pragmatic meaning, while the successful application of betting as an experimental paradigm supports the idea that pragmatic processing of intonation is probabilistic.

Keywords: intonation, pragmatics, perception, Greek

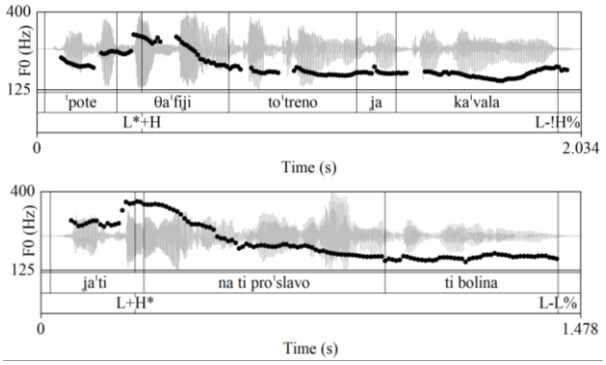
1. INTRODUCTION

A recurrent issue in intonation research is how to model and understand intonational meaning. In some models, intonation is seen as performing basic communicative functions [16], while in others, its main role is said to be the encoding of information structure via accentuation [5]. In the autosegmental-metrical model of intonational phonology [9], intonation is said to perform many roles, including basic functions, such as indicating utterance modality [1], as well as encoding information structure [5], epistemic modality [13], and implicatures [2, 12].

A major disagreement between competing views relates to how intonation contributes to the pragmatic interpretation of utterances. Models like [16] implicitly assume that the entire tune contributes holistically to interpretation. Some models suggest that particular tunes contribute as a whole to pragmatic interpretation, in that tunes can be seen as idioms [8] or as unstructured melodies [9] on some level. These views differ from those expressed in [11] and [12], according to which distinct tonal events each contribute independently to meaning, a position that relies on meaning compositionality.

The issue of compositionality is examined here using Greek wh-questions, the intonation of which has been investigated in the past [1, 2, 3, 7]. These earlier studies show that wh-questions are typically uttered with one of two tunes, autosegmentally represented as L*+H L-!H% (or *rising tune*), and L+H* L-L% (or *flat tune*) [2]. The tunes differ in pitch accent and boundary tone, as illustrated in Figure 1 with two utterances from the present experiment. The acoustic differences between the tunes are well established [2, 7]. Further [2] indicates that the two tunes lead to different interpretations of the questions: questions with the rising tune are treated as primarily information-seeking, while those with the flat tune can be interpreted as non-information seeking. Specifically, [2] show that listeners may interpret the flat tune as carrying implicatures of a negative type; e.g. a question such as "why should I hire Paulina" with the flat tune (Figure 1, bottom) could indicate that the speaker doubts Paulina's credentials. Note that this interpretation is compatible with the main questioning function of the utterance, as the speaker may simultaneously ask for reasons to hire Paulina and indicate that the evidence for doing so is weak.

Figure 1: Waveforms and F0 contours of two wh-questions used as stimuli in the experiment, "when will the train for Kavala leave?" (top), and "why [should I] hire Paulina?" (bottom).



Although this much is established based on previous work [2], issues with this interpretation remain. First, [2] relied on a forced choice task, which did not allow participants to register the possible dual interpretation of the questions. Second, because the

tunes used were the prototypical L*+H L-!H% and L+H* L-L%, it is not possible to ascertain if interpretation was based on separate effects of the pitch accent and following boundary tone, or to the tunes as a whole. Both of these shortcomings are addressed in the present set of experiments.

Participants heard wh-questions and were asked to bet between 0 and 100 euros on the most likely follow-up utterance for each question. The follow-ups provided an explanation of each question's intention as information-seeking or not. This is shown in (1) and (2) with glosses of the two questions in Figure 1 and the two follow-ups (F) for each one of them, one information-seeking (IS) and one non-information-seeking (NIS) follow-up.

(1) Q *When will the train for Kavala leave?*
F-IS *Do you happen to know?*
F-NIS *We've been waiting for too long!*
(2) Q *Why should I hire Paulina?*
F-IS *Is she looking for job?*
F-NIS *She is not qualified.*

It was hypothesized that stimuli with the rising tune would lead participants to bet more heavily on follow-ups indicating that the question was information seeking, and that stimuli with the flat tune would lead participants to bet more heavily on follow-ups indicating that the question was not information seeking. In addition we synthesized tunes to combine early peaks (L+H*) with a final rise (!H%), and late peaks (L*+H) with flat endings (L%).

We hypothesized that these "hybrid" tunes would lead to bets closer to 50, i.e. they would appear more ambiguous. Based on previous accounts [1, 3], it was further hypothesized that participants would be more sensitive to how the tunes ended than to peak location, and thus that rising tunes with early peaks would lead to higher bets for information-related follow-ups than flat-ending tunes with late peaks. Finally, we included controls in the experiment (i.e. questions as they were originally produced). These were included to examine whether cues other than F0 cues (such as changes in speaking rate) could make a contribution to interpretation. We hypothesized that controls would result in higher bets than stimuli, particularly stimuli with the same tune as the controls but a different base (e.g. stimuli with a manipulated rising tune but with a flat tune base; see section 2.2.).

2. METHODS

2.1. Participants

Participants were recruited online, using social media and Prolific (https://prolific.ac/, [10]), a crowd-sourcing platform. Prolific participants were modestly remunerated. All participants were given a chance to win one of eight gift vouchers. The results are based on 190 monolingual participants with no reported speech or hearing problems (150 females, 40 males; $\bar{x}$ age = 31.54, SD = 9.8). Geographically, 40% were from the Athens area, 17.4% from the Peloponnese, 16.8% from Macedonia and 25.8% from other parts of Greece. They were divided as follows: Exp1: 53 (44 female, 9 male; $\bar{x}$ age = 34, SD = 11.2); Exp2: 39 (31 f, 8 m, $\bar{x}$ age = 34.46, SD = 8.7); Exp3: 56 (41 f, 15 m; $\bar{x}$ age = 31.38, SD = 7.9); Exp4: 42 (34 f, 8 m; $\bar{x}$ age = 25.95, SD = 8.8).

2.2. Stimuli

The bases for the stimuli were selected from a larger corpus of Greek wh-questions elicited using a modified Discourse Completion Task, in which participants read short situation descriptions that ended with someone uttering a wh-question [7]. The situations described a background appropriate for either an information-seeking wh-question (i.e. a question with the rising tune), or a non-information seeking version (i.e. a question with the flat tune) [2, 7]. Thirty two questions were selected, four from each of eight talkers (four male), all native speakers of Standard Greek in their early twenties. The selection criteria included how consistent and prototypical the question tunes were, based on previously reported features [2, 7], and on whether they sounded pleasant, natural, and clear. Each talker provided two flat and two rising questions that differed across talkers.

The original questions were manipulated in Praat [4] as described below, following [2, 3, 7].

1. The initial F0 (which can vary between tunes [2]) was scaled to be in the middle of the talker's range for that question.
2. The alignment of the accentual peak was changed to be either (i) *early*, defined as a peak in the middle of the wh-word stressed vowel (henceforth EP) or (ii) *late*, defined as being 20 ms after the onset of the postaccentual vowel (henceforth LP); see [2, 3, 7].
3. The boundary tone was scaled to be either (i) *rising* (defined as being in the middle of the talker's range for that question); or (ii) *flat*. For rising boundaries, the rise began in the middle of the last stressed vowel of the question [3].

These changes were made to the test stimuli only, which were checked for naturalness before being included. The controls (i.e. the original questions) were not altered. The manipulations yielded 128 stimuli (32 Qs × 2 peak alignments × 2 boundary tones), i.e. 4 manipulated versions of each question (Table 1), used in addition to the 32 controls.

Previous experience suggested that presenting participants with both tunes can be confusing and

tiring [7]. To avoid these issues, we split the stimuli and created four experiment versions, each of which was tested with a different group of participants (as shown in section 2.1). For the same reasons, within each version, the stimuli were either rising or flat, and combined with either matched or mismatched controls to give the four versions in Table 2.

Table 1: List of controls and experimental stimuli

Original tune	Manipulated stimuli in AM	Codes
L*+H L-!H%	L*+H L-!H%	LPR
	L+H* L-!H%	EPR
	L*+H **L-L%**	LPF
	L+H* L-L%	EPF
L+H* L-L%	L+H* L-L%	EPF
	L*+H L-L%	LPF
	L*+H L-!H%	LPR
	L+H* **L-!H%**	EPR

In each experiment participants first heard four practice items; these were all original, non-manipulated utterances produced by the same talkers but different from the questions used in the main experiment; they were all either flat or rising to match the controls (see Table 2). Each experiment included 2 controls per talker (16 controls per experiment) and 2 stimuli associated with each control for a total of 48 trials ([16 × 2 = 32 test trials] + [16 controls]).

Table 2: Design of the four experiment versions.

Exp.	Controls	Test Stimuli
1	Late peak & rising (LPR)	Early peak & rising (EPR) Late peak & rising (LPR)
2	Early peak & flat (EPF)	Early peak & flat (EPF) Late peak & flat (LPF)
3	Late peak & rising (LPR)	Early peak & flat (EPF) Late peak & flat (LPF)
4	Early peak & flat (EPF)	Early peak & rising (EPR) Late peak & rising (LPR)

2.3. Procedure

The four versions of the experiment ran in survey mode on PsyToolkit (https://www.psytoolkit.org/, [15]). Participants were instructed to complete the experiment on their laptop, tablet or mobile phone, using headphones (preferred), or if lacking these, in a quiet room. Before the start of the experiment, the following information was provided on separate pages (which participants clicked a button to progress through): (a) the purpose of the experiment (to study tone of voice in Greek), and information about the experimental team; (b) anonymization of data; (c) the participants' right to withdraw from the experiment at any point; (d) instructions about the task. Specifically, the participants were told they would hear a series of questions followed by two written sentences that could be uttered by the speaker as a follow-up to their question. Their task was to decide which follow-up was more likely by placing a bet in a box next to it (see Figure 2). They had to imagine they had 100 euros to bet, and had to bet more than 50 euros on the more likely follow-up, or bet 50 euros on either follow-up if they thought both were equally likely.

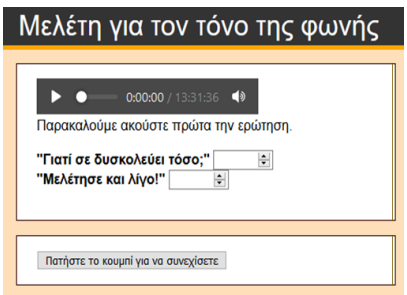
A reminder of the instructions appeared before each trial. Participants clicked on a button to proceed and then saw a screen (Figure 2) where they were prompted to listen to the question and place their bet.

The follow-up utterances were similar to those in (1) and (2). Their order on screen was counterbalanced across trials. To ensure participants always bet on the most likely option, each box was set to accept values of 50 or higher. The 48 trials were divided into 3 blocks of 16; after each block, participants saw a screen that prompted them to take a break if they wished. Each stimulus appeared once. Each talker was heard twice in each block, and stimulus order was pseudo-randomized, so that the two stimuli of each talker in each block were not from the same question. The experiment ended with questions about the participants' linguistic background, and the equipment they used for the experiment.

2.3. Statistical analysis

We ran two linear mixed effects models for each of the four experiments using the *lmer* function of the *lme4* package [6] in R [14]. The results reported here are based on the best fit models; these included stimulus type as fixed factor with three levels, controls (original utterances) and two types of manipulated stimuli that differed per experiment; e.g. in Experiment 1 the levels were EPR and LPR (see Table 2). Participants and items were included as random intercepts [lmer(bets ~ Stimuli + (1|Participants) + (1|Items), data = Exp, REML = FALSE). These models performed better than the corresponding null models according to the likelihood ratio test [6, 14], [Exp1: $x^2(2) = 81.828$, $p < 0.001$, Exp2: $x^2(2) = 7.605$, $p < 0.05$, Exp3: $x^2(2) = 484{,}378$, $p < 0.001$, Exp4: $x^2(2) = 165.187$, $p < 0.001$].

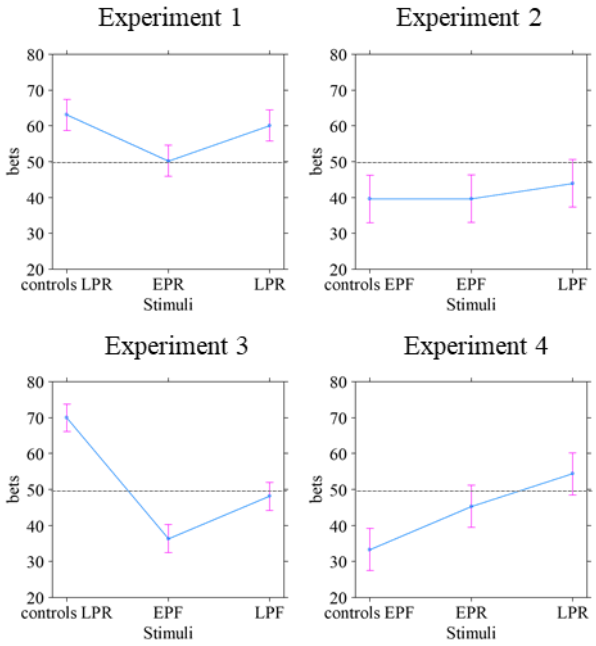
Figure 2: The screen seen by participants at each trial.



3. RESULTS

For reasons of space, we report only on comparisons of interest (see also Figure 3). Experiment 1 yielded 2428 responses. Participants placed significantly higher bets on information-seeking follow-ups after LPR stimuli and controls than after EPR stimuli [for LPR stimuli, est. = 9.875, SE = 1.475, $p < 0.001$; for LPR controls, est. = 2.962, SE= 1.473, $p < 0.05$]. Bets were also significantly higher after LPR controls than LPR stimuli [est. = 2.962, SE = 1.473, $p < 0.05$].

Figure 3: Effect plots of the stimuli types in the linear mixed effects models. The y-axis represents the value of bets for information-seeking follow-ups. The broken grey line represents bets at chance level.



Experiment 2 yielded a total of 1869 responses. Participants placed significantly higher bets on information-seeking follow-ups after LPF than EPF stimuli [est. = 4.224, SE = 1.795, $p < 0.02$] and EPF controls [est. = 4.357, SE = 1.796, $p < 0.02$]. There was no significant difference between EPF controls and EPF stimuli [est. = 0.132, SE = 1.796, $p > 0.05$].

Experiment 3 yielded 2654 responses. Participants placed significantly higher bets on information-seeking follow-ups after LPR controls relative to both EPF and LPF stimuli [for EPF, est. = 33.605, SE = 1.476, $p < 0.001$; for LPF, est. = 21.831, SE = 14.77, $p < 0.001$]. Bets were higher after LPF than EPF stimuli [est. = 11.774, SE = 1.474, $p < 0.001$].

Experiment 4 yielded 2008 responses. Participants placed significantly higher bets on LPR stimuli relative to both EPR stimuli [est. = 12.030, SE = 1.613, $p < 0.001$] and EPF controls [est. = 21.101, SE = 1.612, $p < 0.001$]. Participants also placed significantly higher bets on EPR than EPF controls [est. = 9.071, SE = 1.612, $p < 0.001$].

4. DISCUSSION AND CONCLUSION

The aim of these experiments was to shed light on the compositionality of intonational meaning. Listeners heard and bet on both prototypical tunes, L*+H L-!H% (*late peak & rising*) and L+H* L-L% (*early peak & flat*), and their combinations L*+H L-L% (*late peak & flat*) and L+H* L-!H% (*early peak & rising*).

The results support both the overall understanding of the pragmatics of the two tunes [cf. 2], and compositionality, as both peaks and boundaries affected responses and made independent contributions to bets. Overall, stimuli with late peaks lead to significantly higher bets in favour of information-seeking follow-ups as compared to stimuli with early peaks; the effect was present independently of the boundary tone. Similarly, rising tunes lead to bets over 50 for information-seeking follow-ups, while flat tunes consistently led to bets below 50, independently of the pitch accent. As hypothesized, bets were highest for information-seeking follow-ups when final rises were combined with late peaks. It could be argued that the contribution of the peaks was exaggerated because there was no variation in the boundary tone within each experiment. Although this applies to Exp1 and Exp2, it does not apply to Exp3 and Exp4 where controls and stimuli did not match for boundary tone. Yet in those experiments as well, peak position was significant (cf. Exp2 and Exp3, and Exp1 and Exp4).

The independent contribution of the peak and boundary is supported by the results with hybrid tunes: as expected, these led to bets closer to chance (cf. EPR and LPR in Exp1). It is possible that these tunes were judged non-representative and thus more confusing; however, they are attested in Greek [2]. Additional research in preparation using betting with masking should shed further light onto this issue.

Finally, the results from controls highlight the role of cues other than F0 (such as segmental timing [7]). Such cues clearly play a part, in that controls led to significantly stronger bets, either in favour of information-seeking follow-ups (Exp1, Exp3), or non-information-seeking follow-ups (Exp4). The controls were also used as the base for the stimuli of each experiment; when the base and tune matched (Exp1, Exp2), betting trends were more consistent than when there was a mismatch (Exp3, Exp4), a difference that further supports the role of cues beyond F0 in interpreting intonation.

In conclusion, the findings of this study support the notion that accents and edge tones contribute independently to pragmatic meaning. They also showcase the suitability of the betting paradigm when seeking interpretable results bearing on compositionality in intonation.

5. REFERENCES

[1] Arvaniti, A., Baltazani, M. 2005. Intonational analysis and prosodic annotation of Greek spoken corpora. In: Jun, S.-A. (ed), *Prosodic Typology: The Phonology of Intonation and Phrasing*. Oxford University Press, 84–117.

[2] Baltazani, M., Gryllia, Arvaniti, A. 2019. The intonation and pragmatics of Greek wh-questions. *Language and Speech*. https://doi.org/10.1177/0023830918823236

[3] Arvaniti, A., Ladd, D. R. 2009. Greek wh-questions and the phonology of intonation. *Phonology* 26: 43–74.

[4] Boersma, P., Weenink, D. 2018. Praat: doing phonetics by computer [Computer program]. Version 6.043, retrieved 8 September 2018 from http://www.fon.hum.uva.nl/praat/.

[5] Büring, D. 2016. *Intonation and Meaning*. Oxford: Oxford University Press.

[6] Douglas, B., Maechler, M., Bolker, B., Walker, S. 2015. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software* 67(1), 1–48.

[7] Gryllia, S., Baltazani, M., Arvaniti, A. 2018. The role of pragmatics and politeness in explaining prosodic variability. *Proceedings of Speech Prosody 2018.*

[8] Gussenhoven, C. 2004. *The Phonology of Tone and Intonation*. Cambridge University Press.

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

[10] Palan, S., Schitter, C. 2018. Prolific.ac – A subject pool for online experiments. *Journal of Behavioral and Experimental Finance* 17, 22–27.

[11] Pierrehumbert, J. 1980. *The phonology and phonetics of English intonation*. PhD thesis, MIT.

[12] Pierrehumbert, J. B., Hirschberg, J. 1990. The meaning of intonational contours in the interpretation of discourse. In: Cohen, P. R., Morgan, J. L., Pollack, M. E. (eds), *Intentions in Communication*. The MIT Press, 271–311.

[13] Prieto, P., Borràs-Comes, J. 2018. Question intonation contours as dynamic epistemic operators. *Natural Language and Linguistic Theory* 36(2), 563–586

[14] R: A Language and Environment for Statistical Computing.

[15] Stoet, G. (2010). PsyToolkit - A software package for programming psychological experiments using Linux. Behavior Research Methods, 42(4), 1096–1104.

[16] Xu, Y. 2005. Speech melody as articulatorily implemented communicative functions. *Speech Communication* 46(3-4): 220–251.