

Prosody, emotions, and... ‘whatever’

Stefan Benus¹, Agustín Gravano², Julia Hirschberg²

¹Department of Cognitive and Linguistic Sciences, Brown University, Providence, RI, USA

²Department of Computer Science, Columbia University, New York, NY, USA

sb513@nyu.edu, agus@cs.columbia.edu, julia@cs.columbia.edu

Abstract

We examine the role of prosody in cueing a scale of negative meanings associated with the use of *whatever*. The analysis of a corpus of elicited examples shows that the more negative the token, the more likely it is to have an additional pitch accent, extended duration, and expanded pitch range on the first syllable. These findings are analyzed as a link between pragmatic meaning and the strength of the prosodic boundary between the first two syllables (*what#ever*). The results of perception experiments show that the prosody of *whatever* itself is a systematic cue for the degree of negative connotation associated with the utterance in which *whatever* occurs. Potential applications of this result for spoken dialogue systems and synthesis of emotional speech are discussed.

Index Terms: prosody, pragmatic meaning, emotional speech.

1. Introduction

Like other pragmatic markers, *whatever* has a range of pragmatic interpretations. If used as a modifier, *whatever* may imply ignorance or indifference (or both) on the part of the speaker [1] and may function as an indiscriminate or quodlibetic free-choice marker similar to *any* [2]. For example, the indifference / ignorance meaning is illustrated in utterances such as “Pick whatever apple you want” [1]. There is a presupposition of ignorance, since the speaker doesn’t know which apple the hearer wants, and also indifference, since the speaker does not seem to care which apple will be picked.

However, relatively recently, *whatever* has developed a pragmatic meaning signaling a continuum of attitudes between neutral and negative and, in some cases, a wish to terminate the current discourse segment [3]. For example, B’s response in (i) below can signal neutral indifferent attitude: “Yeah, I don’t care, do it whenever you can”, *or* a negative attitude: “I hate your laziness; let’s leave it at that”.

- (i) A: Can I do it tomorrow?
B: Whatever

From a corpus of overheard and transcribed short conversations, [3] identified three main categories of non-modifier *whatever* usage; these are exemplified below: filler (ii), neutral marker (iii), and negative evaluation marker of the proposition, the interlocutor, or both (iv):

- (ii) I don’t wanna waste my time buying a prom dress or whatever.
(iii) A: Hey Ritchie, you want these over here?
B: Yeah, whatever, just put them down.
(iv) So she ordered all this stuff and two days ago she changed her mind. I was like, whatever.

It was suggested that a more negative meaning is signaled by occurrences of *whatever* that have longer duration, greater pitch excursion, and constitute a separate intonational phrase. However, testing these observations was not possible in the absence of recorded speech data.

We hypothesize that the valence scale of pragmatic meanings of *whatever* has developed from the ‘indifference’ conveyed by its modifier meaning amplified by prosodic variation. Prosody is an important cue for signaling emotional and attitudinal meanings, but the diachronic process of pejoration [4] has not yet been linked to prosody. While prosodic features such as higher F0, intensity, and speaking rate have been shown to correlate with the degree or emotional involvement (activation) [5], there has been less success in relating prosodic variation to emotional valence differences [6]. The case of *whatever* provides an opportunity to investigate the correlation between emotional meaning and prosody realized on a single word, and thus keeping the segmental material constant.

In this study, we report results of production and perception experiments that test the correlation between multiple prosodic features and the pragmatic meaning of *whatever*. Sections 2 and 3 describe the methodology and results of the production and perception experiments, respectively. Section 4 discusses the main results and their potential applications.

2. Production study of *whatever*

2.1. Corpus

Use of *whatever* as a negative evaluation marker requires emotional involvement that is difficult to simulate in interviews and other laboratory collection techniques of spontaneous speech. Existing corpora with a wide variety of *whatever* uses (e.g. *The Jerry Springer* show) are typically poor in sound quality. To achieve a balance in the breadth of the data and its sound quality, we collected instances of *whatever* using acted speech. We used 5 conversations from the corpus of naturally occurring overheard and transcribed conversations [3] that included 10 instances of *whatever*, similar to examples (ii)-(iv) above and (v) below. These 10 tokens were rated on a scale from neutral (1) to most negative (4) by the first author and the ratings were confirmed by two other independent raters. In designing the scale we followed the observations in [3] that *whatever* does not convey positive meanings and that it may convey up to three degrees of non-neutral negative meaning. The 10 tokens were divided as follows: three tokens in both NEUTRAL (1) and VERY NEGATIVE (4) categories and two tokens in both SLIGHTLY NEGATIVE (2) and NEGATIVE (3) categories. Conversation (v) below is an example of the stimuli presented to subjects including the comments in the brackets; note that this stimuli contains three instances of *whatever*.

- (v) A family situation, an older sibling talking to a younger one:
 - A: Mom says you aren't studying enough or practicing your violin.
 - B: OK, whatever. (accepting the reality)
 - A: But you really should start doing it.
 - B: Whatever. (I don't like what you are suggesting, I am getting angry)
 - A: So when are you going to practice? Mom says that you need to everyday.
 - B: Whatever. (I've got enough, stop bugging me)

We asked 12 subjects, all females aged 20-35, to read the transcripts of the 5 conversations. Before the recording, the subjects could study the five situations and ask questions. During this stage, the experimenter made sure that subjects' understanding of the situations was consistent with the intended meanings of *whatever*. Subjects were instructed to read them as naturally as possible, imagining themselves in those situations. Once they felt comfortable with the situations, they were recorded in a sound-proof booth with the experimenter present. The recordings were then digitized at 22 kHz. This procedure yielded 120 tokens of *whatever*.

2.2. Data labeling and extraction

All tokens were transcribed, hand-aligned with the speech signal, and analyzed using Praat [7]. We labeled syllable boundaries and the duration of the coronal closure for /t/. Additionally, all tokens were ToBI labeled [8] for pitch accents, phrase accents and boundary tones. We extracted maximum and minimum F0 value for each syllable and the whole *whatever*, and calculated the pitch range associated with each syllable and token.

2.3. Results

Our production study indicated the most salient prosodic factor in cuing the pragmatic meaning of *whatever* to be the number of pitch accents associated with the word itself. More negative meanings tend to be correlated with tokens uttered with more pitch accents, Pearson's $r = 0.657$, $p < 0.001$ (Fig. 1). All but one token with two pitch accents has a negative meaning. Moreover, most doubly accented tokens resulted from the stimuli with negative meanings. Out of a total of 36 doubly accented tokens, 67% fall in the VERY NEGATIVE (4) category and 30% in the NEGATIVE (3) category. At the other end of the pragmatic spectrum, tokens in the NEUTRAL (1) category differ from the SLIGHTLY NEGATIVE (2) cases primarily by the percentage of the tokens with no pitch accent at all.

In terms of pitch accent type, negative meaning tends to be signaled by a sharply rising pitch accent (L+H*) while neutral meaning is signaled by a H+!H* accent; Pearson's chi-square test, $\chi^2 = 50.18$, $df = 6$, $p \approx 0$ (Fig. 1). No significant difference was observed for the phrase and boundary tones, which tended to be L-L% for all productions.

There was also a significant effect of meaning on duration: the more negative the meaning, the longer the total duration of *whatever*, as well as the duration of each syllable separately (Table 1). The most robust results were observed for the first and third syllable durations; $r = 0.575$, $p < 0.001$ and $r = 0.566$, $p < 0.001$ respectively.

In a subset of the data (N = 30) the duration of the coronal closure was labeled. A one-way analysis of variance (ANOVA) revealed that the longer closure duration correlated with more negative meaning. A significant effect was found

when the data were pooled into VERY NEGATIVE (4) vs. OTHER (1-3), $F(1, 28) = 6.87$, $p = 0.006$.

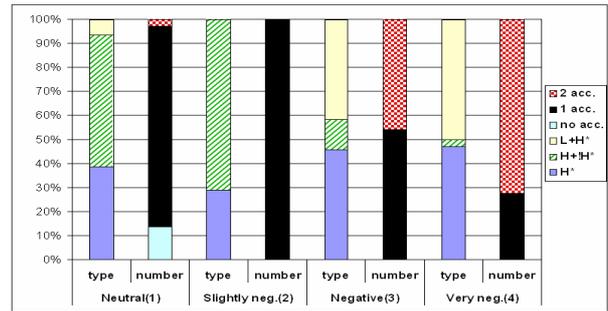


Figure 1: Number and type of pitch accents as a function of meaning

Pragmatic meaning appears to affect pitch range in a similar manner: the more negative productions of *whatever* have a larger pitch range (Table 1). Although the pitch range in each syllable and in total was affected by meaning significantly, the most robust result showed for the first syllable ($r = 0.477$, $p < 0.001$).

Table 1. Duration (ms) and pitch range (Hz) as a function of meaning.

Meaning	Syll-1		Syll-2		Syll-3		Total	
	Dur	Range	Dur	Range	Dur	Range	Dur	Range
1	106	53	101	28	151	29	357	129
2	119	56	104	27	168	29	392	145
3	136	53	114	54	197	52	446	139
4	233	131	134	64	236	72	603	207

In summary, our production experiment showed that, as *whatever's* interpretation becomes more negative, the first syllable is more likely to carry a pitch accent, more likely to be lengthened and more likely to be uttered in an expanded pitch range. Thus, prosodic cues offer systematic predictions as to the degree of negativity of *whatever*.

3. Perception studies of *whatever*

3.1. Procedure

In order to investigate the prosodic variation of *whatever* in naturally occurring speech, we analyzed an on-line speech corpus HUB-5 ([9]), consisting of 200 telephone conversations between friends. There are 140 tokens of *whatever* with non-literal meanings. The vast majority of these tokens are fillers, like example (ii) in Section 1; however, there is fairly good coverage of other meanings. For the perception experiment, we selected 12 tokens from this corpus in an effort to cover different degrees of negativity (1-4) equally. To increase the variety and size of our stimuli we also analyzed several hours of daytime talk shows such as *Jerry Springer* and *Maury*, but due to excessive noise, only one token was added to the stimuli list. Finally, we included one token from a sociolinguistic interview conducted for a different study, which made the total number of tokens 14.

Separate perception tests were administered to three different groups of subjects. In the first (PROSODY-ONLY), the subjects were presented with the transcript of 14 identical frame dyads (vi) and each dyad was followed by a sound file with a token of *whatever* which had been excised from its original context and presented through a loudspeaker.

Subjects were instructed to imagine themselves in this frame situation and asked to rate, on a scale between 1 and 7, how positive (1) or negative (7) they perceived B's response.

- (vi) A: Do you wanna get some Chinese for dinner?
B: Whatever.

Hence, in this study, subjects made their judgments based only on prosody, with no other cues from the original context since they only had a fake context and it was always the same for all 14 tokens. In the second study (CONTEXT-SPEECH), subjects listened to the original speech files containing the tokens of *whatever* together with the context in which it occurred. They were asked to rate the polarity of *whatever* on the same scale as in the first study. In the third study (CONTEXT-TRANSCRIPT), subject rated *whatevers* based on transcriptions of the original contexts alone. Subjects for all three tests were undergraduate students, ages 19-21. 38 subjects participated in the first, 19 in the second, and 19 in the third test.

3.2. Results

3.2.1. 'Whatever' in a prosody-only condition

We first asked whether the prosodic features found to correlate with negativity in acted speech would function similarly in naturally occurring data. The production results are supported even in this very small corpus. Tokens rated as more negative tended to have longer coronal closure and first syllable, and greater pitch range in the first syllable than more neutral tokens. Due to the small number of tokens, only the coronal closure achieved significance at 0.05, $r = 0.56$, $p = 0.037$, and a tendency was observed between meaning and first syllable duration, $r = 0.46$, $p = 0.097$. Table 2 shows the prosodic features of the 14 tokens arranged from the most to the least negative.

Table 2. Prosodic features and token ratings in a fixed context.

Token	rating	/t/-clo.	dur-1-syll	range 1-syll	accent
1	5.79	86	295	85	L+H* !H*
2	5.03	15	93	28	H+!H*
3	5.00	46	126	30	H*
4	4.79	23	103	8	H+!H*
5	4.47	28	121	13	H*
6	4.34	12	116	20	H*
7	4.29	10	101	19	H*
8	4.24	17	111	2	H+!H*
9	4.18	32	120	7	H*
10	4.08	38	195	12	H*
11	3.95	22	83	45	L+H*
12	3.71	23	147	39	H*
13	3.58	23	130	43	H*
14	2.16	16	93	22	L+H*

Only one token of the 14 (token 1) was uttered with 2 pitch accents, and this token was judged as the most negative by a significant margin (0.8). All other tokens were uttered with a single pitch accent. In terms of accent type, token 1 had a rising L+H* accent on the first syllable, shown in the production study to correlate with negative meaning. As seen in Table 2, all other prosodic features of token 1 have the greatest values compared to all other tokens. This observation suggests that the prosodic features identified in the acted speech production study described in Section 2.3 are also

used to signal negative *whatever* in naturally occurring speech.

For the remainder of the data, the differences both in terms of prosody as well as pragmatic meaning are less clear. For example, in addition to token 1, the L+H* accent occurs on two tokens that were rated among the *least* negative (11, 14). This seems to contradict the correlation of this accent with negative meaning observed in Section 2. Note, however, that these two tokens have also the lowest values for duration of first syllable, suggesting that it is the combination of duration and pitch accent features which signal negativity.

From ratings of other tokens, it is clear that other factors must be at play in signaling negative attitude. For example, token 2 is rated as the second most negative although values for all prosodic features are quite low and the pitch accent is H+!H* (an accent type linked to neutral meaning in the production experiment). This token, however, has a very flat F0 contour and extremely long third syllable (Fig. 3). Uninterested and bored attitude is usually signaled with a very flat F0 contour and prolongation. Tokens 4 and 5, rated as fourth and fifth most negative, respectively, have a similar plateau F0 contour. Such 'plateau' contours in a narrow pitch range often signal boredom and lack of interest, which is a fairly negative emotion. Hence, we may hypothesize along with [6] that plateau pitch contour signals negative emotions and thus, in our data, cues a negative interpretation of *whatever*.

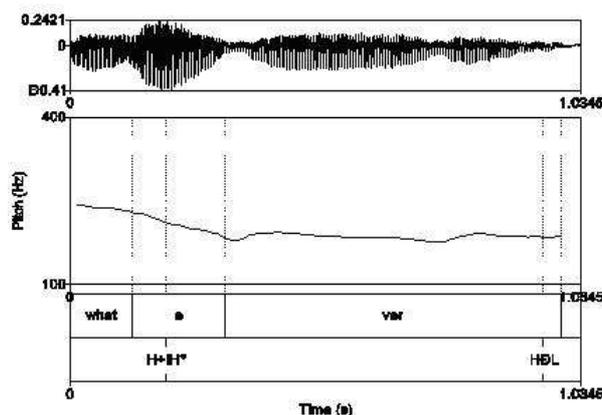


Figure 3: Sound wave (top panel) and F0 (middle panel) of Token 2

3.2.2. 'Whatever' in context

Recall that two groups of subjects judged *whatever* together with the contexts in which it occurred. In the first group, subjects heard the situations (CONTEXT-SPEECH); in the second, subjects read the transcripts of these situations (CONTEXT-TRANSCRIPT). We were interested in how judgments in these two conditions would correlate with the perception of *whatever* when we controlled for the effect of context (PROSODY-ONLY, see Section 3.2.1). We found that subjects' ratings in the CONTEXT-TRANSCRIPT condition strongly correlated with the ratings in the CONTEXT-SPEECH condition ($r = 0.8$, $p < 0.001$). Interestingly, a strong positive correlation was also found between PROSODY-ONLY and CONTEXT-TRANSCRIPT ($r = 0.67$, $p = 0.009$), and between PROSODY-ONLY and CONTEXT-SPEECH ($r = 0.595$, $p = 0.025$). Mean subjects' ratings in the three conditions are shown in Fig. 4.

The strong correlations between the perceived degree of negativity in the PROSODY-ONLY and CONTEXT-TRANSCRIPT

conditions, and in the PROSODY-ONLY and CONTEXT-SPEECH conditions, seem to suggest that the prosody of *whatever* itself carries a substantial amount of information about the negativity of the original situation.

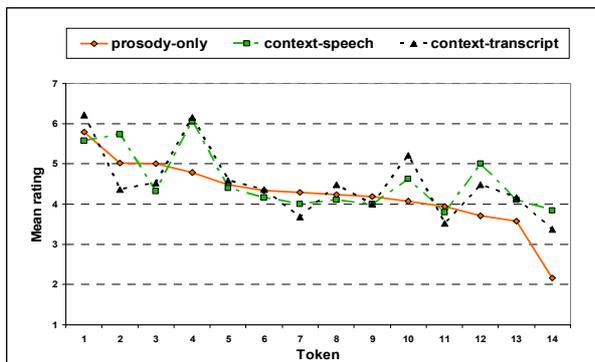


Figure 4: Mean ratings in three perception tests

Several additional observations can be made from these perception judgments. An ANOVA revealed that subjects perceived *whatever* on average as significantly more negative in context than without it (mean ratings of PROSODY-ONLY: 4.26, CONTEXT-SPEECH: 4.55; $F(1, 795) = 7.03$, $p < 0.01$). This finding is corroborated by examining the mean ratings for PROSODY-ONLY and CONTEXT-SPEECH separately for each token. Two-sided t-tests showed that out of 5 tokens with significant differences between the judgments (at $p < 0.05$), 4 were perceived as more negative in context than in the PROSODY-ONLY condition. Additionally, the presence of the original context made the subjects' judgments on average slightly more uniform (mean standard deviation of PROSODY-ONLY: 1.31, CONTEXT-SPEECH: 1.05).

The increased negativity in the context condition may stem from the fact that the fixed context in PROSODY-ONLY is assumed to be neutral. In fact, some discrepancies between ratings in context and prosody-only conditions can be explained by the use of charged language not directly related to *whatever*. Note that even though subjects were explicitly instructed to rate the meaning of *whatever* only, in their judgments they were likely to have been influenced by the overall meaning of the situation and general negative or positive emotion conveyed by the speaker also. The addition of contextual cues makes the pragmatic meaning of *whatever* more negative especially in tokens 4, 10, 12, and 14. One of these appeared in a dialogue from *Jerry Springer* in which a woman used *whatever* as a reply to being called "little tramp", and another in a monologue in which "shit" was used twice.

In summary, the perception experiments showed that the prosody of *whatever* is a fairly reliable predictor for the degree of negativity associated with the situation in which it occurred, but, in some cases, lexical contextual cues affect the perception of negativity as well.

4. Discussion

Results of production and perception studies show that prosody provides systematic cues to the pragmatic meaning of *whatever*. The salient prosodic cues include the presence of a pitch accent on the first syllable, extended duration and pitch range on this syllable, rising pitch accent, and longer duration of coronal closure at the boundary of this syllable. Together, these cues are consistent with the hypothesis that the degree of pragmatic negativity correlates with the strength of the

prosodic boundary between the first two syllables of *whatever*. Additionally, support was found for the correlation between negative meaning and flat (plateau) pitch contour combined with final lengthening. Data also indicated that prosodic cues may sometimes be overridden by situational context.

The ability to detect negative emotion is widely recognized as critical in applications such as spoken dialogue systems, where determining user satisfaction is crucial. Since it is plausible that an emotionally ambiguous word such as *whatever* is uttered in diverse situations by users of such a system, an incorrect evaluation of its degree of negativity might dramatically increase the chances of system/user miscommunication. Our results suggest that the degree of negativity of a whole utterance or situation may be predicted solely from the prosody of individual *hot-spot* words such as *whatever*. Future research should further study this hypothesis (and possibly extend the analysis to other common potentially trigger words such as *fine*, *sure*, *please*, or *yeah*), since if it were true, it would then be possible to use the prosody of these hot-spot words in facilitating the rather complex task of predicting the degree of negativity of the whole situation or utterance.

Additionally, synthesis of emotional speech could benefit from our findings. If the valence of emotion to be expressed is known, synthesizing the hot-spot word with the appropriate prosodic features may increase the chances for perceiving the message as intended.

An open question remains as to whether prosody can signal a continuous scale of negative attitudes, or if prosodic cues are used for binary disambiguation between neutral and negative attitude. The predictive power of discrete features (pitch accent) and the bimodal distribution of gradient features (duration, pitch range) suggest that a prosody-pragmatics mapping may provide a binary division between neutral and negative, while additional variation may be accounted for by individual implementation of this mapping.

5. References

- [1] Von Fintel, K., "Whatever". Proceedings from SALT 10: 27-40, 2001.
- [2] Horn, L., "Any and (-)ever: Free choice and free relatives". Proceedings of IATL 15, 1999.
- [3] Blake, R., Bakht-Rofheart, M., Benus, S., Cooper, S., Josey, M., and Solyom, E., "I have three words for you...": Whatever as a discourse marker". Presented at NWAV 27, 1999.
- [4] Hock, H.H., and Joseph, B. D., "Language History, Language Change, and Language Relationship: An Introduction to Historical and Comparative Linguistics", Mouton de Gruyter, Berlin-New York, 1996.
- [5] Schröder, M., Cowie, R., Douglas-Cowie, E., Westerdijk, M., Gielen, S., "Acoustic Correlates of Emotion Dimensions in View of Speech Synthesis", Eurospeech-2001, Aalborg, 87-90.
- [6] Liscombe, J., Venditti, J., and Hirschberg, J. "Classifying subject ratings of emotional speech using acoustic features", Proceedings of Eurospeech, 2003.
- [7] Boersma, P., Weenink, D. "Praat: Doing phonetics by computer". [Computer program] <http://www.praat.org>.
- [8] Beckman, M. E., Hirschberg, J. 1994. "The ToBI annotation conventions". Ohio State University, 1994.
- [9] Graff, D., Martin, A., and Miller, D., "HUB5 English Evaluation". Linguistic Data Consortium, Philadelphia, 2002/1998.