EVALUATING DIFFERENT RATING SCALES FOR OBTAINING JUDGMENTS OF SYLLABLE PROMINENCE FROM NAÏVE LISTENERS

Denis Arnold[a], *Petra Wagner*[b] & *Bernd Möbius*[c]

[a]Language and Speech Communication, University of Bonn, Germany;
[b]Faculty of Linguistics and Literature, Bielefeld University, Germany;
[c]Department of Computational Linguistics and Phonetics, Saarland University, Germany
dar@sk.uni-bonn.de; petra.wagner@uni-bielefeld.de; moebius@coli.uni-saarland.de

ABSTRACT

Streefkerk [6] defines prominence as the perceptually outstanding parts in spoken language. An optimal rating scale for syllable prominence has not been found yet. This paper evaluates a 4-point, an 11-point, a 31-point, and a continuous scale for the rating of syllable prominence and gives support for scales using a higher number of levels. Priming effects found by Arnold, et al. [1], could only be replicated using the 31-point scale.

Keywords: prosody, syllable prominence, rating scales, perception, methodology

1. INTRODUCTION

Despite general agreement on the term *prominence* denoting the degree of perceptual salience, there have been numerous ways of capturing it, e.g. by using different rating scales. It is possible that – among other things - conflicting findings on how prominence is linked to acoustics are the result of the different ways prominence has been measured in those studies. Our investigation aims at determining the best strategy of obtaining prominence ratings by listeners.

Jensen and Tøndering [4] compared 2-point, 4-point and 31-point scales for the rating of word-level prominence. They found that ratings obtained with the different scales are very similar. They argued that a 31-point scale is more difficult to handle for non-expert listeners than a 4-point scale. Therefore, they concluded that a 4-point scale is optimal. Grover, et al. [3] reported that they found more reliable results when using a 10-point scale to rate syllable prominence compared to a 4-point scale. Eriksson, et al. [2] used continuous sliders to obtain judgements of syllable prominence from their subjects. The same approach was used by Windmann, et al. [7]. They calculated the prominence value from the percentage of the range of the sliders. Arnold, et al. [1] used a 31-point scale with sliders to identify the influence of priming on the perception of syllable prominence in German.

The purpose of this paper is to investigate the inconsistency between the findings of [3] and [4]. We wanted to compare the usage of sliders without number labels to other scales. Another goal was to find out whether all evaluated scales are able to detect the effects of priming found in [1].

2. EXPERIMENT

216 subjects were asked to rate the syllable prominence of 15 German sentences based on different rating scales. We chose to evaluate a 4-point, an 11-point, a 31-point, and a continuous scale using sliders, by means of a graphical user interface based on java-swing.

2.1. Design

The experimental design comprises 72 groups of subjects. Each subject rated the stimuli using two different rating scales. First, all stimuli where rated using the first rating scale and then all stimuli where rated using another scale. To control order effects we combined all four scales in all possible orders, which resulted in 12 subject groups. To test the effect of priming we needed to duplicate the number of groups. We also manipulated the instructions given to the subjects, aiming for three differed levels of accuracy, resulting in 72 groups.

2.2. Manipulation of instructions

The instructions varied in how often the subjects should listen to the signals. Whereas it is crucial for priming that the subjects listen to the signal at most twice, listening to the signal more often might increase the quality of the judgement. We instructed the first group to listen to the signal at most twice. We asked the second group to listen to

the signal several times in order to perceive fine differences between the syllables. The subjects in the third group were told that they could listen to the signal again if they liked.

2.3. Rating scales

All rating scales were implemented as J-Sliders with a length of 300 pixels. We designed a special class to hide the slider knob until a syllable was rated for the first time to avoid a possible influence of the initial slider position on the subjects' ratings.

The 4-point scale had four tick marks and the 11-point scale had eleven tick marks with labels indicating the values. The 31-point scale had a mixture of tick marks without labels and tick marks indicating the steps from 0 to 30 in increments of five.

For these three scales the prominence value was computed from the slider position with the standard methods of the J-Slider Class.

The continuous scale had two labels indicating the maximum and minimum. Internally the scale had 300 steps using the maximal resolution.

2.4. Priming

Different studies have shown that perception of syllable prominence is not purely bottom-up driven but also guided by linguistic knowledge and expectations.

Arnold, et al. [1] used the priming paradigm to directly manipulate the expectations about the prominence patterns of sentences. Subjects where manipulated to associate a certain syntactic and semantic structure with a specific prominence pattern. The manipulation caused a significant difference in the ratings of syllable prominence for a sentence if it had the same syntactic and a similar semantic structure as the priming sentences.

The study used a 31-point scale to obtain judgements of syllable prominence. In the present study we ask whether the same results can be obtained if a different rating scale is used. We used the following set of priming sentences from [1] (Italic typesetting and underline indicates prominent syllable).

Group 1:
test sentence:
Die *jun*ge Frau geht in das rote Haus.

priming sentences:
Der alte Mann stieg in den vollen Bus.
Das kleine Kind ging in das kleine Haus.
Die alte Frau steigt in den leeren Bus.
Der junge Mann geht in das gelbe Haus.

Group 2:
test sentence:
Die *jun*ge Frau geht in das rote Haus.

priming sentences:
Der *al*te Mann stieg in den vollen Bus.
Das *klei*ne Kind ging in das kleine Haus.
Die *al*te Frau steigt in den leeren Bus.
Der *jun*ge Mann geht in das gelbe Haus.

The expectation is that the rating on the syllable "*jun*" in the test sentence differs significantly between the two groups.

2.5. Speech material

We used 10 sentences comprising 3 to 10 syllables and different prominence patterns. For the priming we used the test sentences and priming sentences described in Arnold, et al. [1]. There are four priming sentences and one test sentence for each priming condition. Thus, every subject rated 15 stimuli two times each.

The speech material was spoken by a trained speaker and was not modified. The stimuli were recorded in a sound-treated studio and stored as 16-bit, 44.1 kHz wave files.

2.6. Rating experiment

The experiment was carried out by means of Java coded software. All instructions were presented on the computer screen. The stimuli were presented via headphones and the subjects were asked to judge the prominence of each syllable using sliders on the GUI. The orthographic representation of each syllable was shown above the corresponding slider. The subjects had to rate all syllables of a sentence before proceeding. They had the opportunity to listen to the signals again if they wished to, using a button on the graphical user interface.

3. RESULTS

3.1. Manipulation of the instructions

We found that the manipulation of the instructions successfully affected the number of playbacks and time consumption of the subjects.

The difference between the first and second group is always significant. The results of the third group were between those of the other two groups, sometimes closer to the first and sometimes closer to the second group.

3.2. Utilization of the range of the scales

Subjects used all levels of all scales and the distributions are quite similar between the different scales. Thus, subjects seemed to be able to make full use of a "continuous" scale using 300 steps. Values that function as an anchor on the slider received higher scores in the 31-point and the continuous scales.

3.3. Extreme results and deviations

We found that the 11-point, 31-point and continuous scales had a smaller range of median values compared to the 4-point scale. Syllables that are dominantly prominent were likely to receive extreme rating values independent of the actual number of levels of the scale.

Figure 1: Box plot for the ratings of a sentence obtained with the 4-point scale.

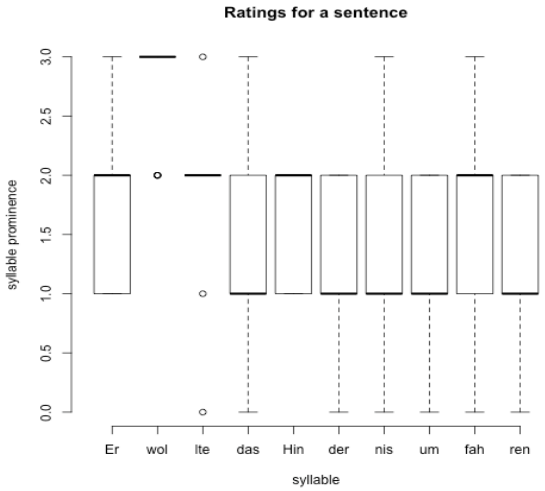
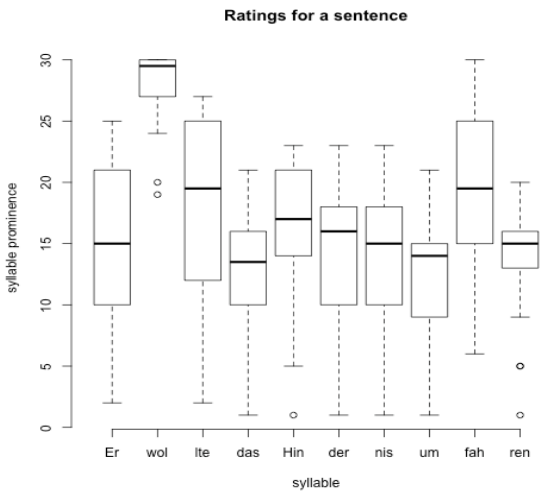


Figure 2: Box plot for the ratings of a sentence obtained with the 31-point scale.



An interesting aspect of using scales with more levels can be seen in figures 1 and 2. While the 4-point scale has a limited range of possible deviations for each rated syllable, we observed that the deviation on the scales with more levels differed more between the syllables. This means that subjects agreed more on the prominence of certain syllables than on others. Why inter-subject agreement varies across syllables remains an open question due to the restricted nature of the test material in the present study.

3.4. Ratings and acoustic correlates

Important aspects of the research on prominence of linguistic units are the correlations with acoustic features such as f0, duration and intensity. We computed linear regression analyses by means of R [5] using f0, intensity and duration as predictor variables for prominence. We found explained variances (r^2) between .18 and .40 and a higher contribution of intensity and duration than of f0.

The correlations between the ratings and the acoustic correlates are in line with findings from other studies.

The explained variance r^2 varied from .18 to .35 for the 4-point scale, from .22 and .40 for the 11-point scale, from .20 and .32 for the 31-point scale and from .27 and .37 for the continuous scale.

Depending on the instruction and priming condition each scale received the highest explained variance in the particular setting at least once.

3.5. Playbacks and time consumption as an indicator of rater effort

It is crucial for the success of a study that the rating process is easy for the subject. Jensen and Tøndering [4] used the time it takes a subject to complete the task and the number of playbacks as an indicator for the difficulty of using a given rating scale.

In the present study the differences in time consumption between the 11-point scale and 31-point scale were not significant. No differences in the number of playbacks were significant.

Table 1: Mean time consumption and mean number of playbacks for the different scales.

	4-point	11-point	31-points	continuous
Time [sec]	25.76	28.83	29.92	27.82
# playbacks	1,13	1,15	1,21	1,11

3.6. Correlation of the results using the different scales

We found correlations ranging from .74 to .85 between the ratings for different scales, instructions and priming conditions, depending on the scales and on the manipulation of the instructions.

3.7. Success of detecting effects of priming

The study of Arnold, et al. [1] used a 31-point scale. The question is whether the effects of priming could have been shown with the other scales evaluated in this paper.

We used one set of priming stimuli described in Arnold, et al. [1] and tried to reproduce the results. We compared the differences D_n defined by equation 1 (from Arnold, et al. [1]), where P_n is the prominence of the syllable on position n in the utterance.

$$D_n = \frac{2P_n - P_{n-1} - P_{n-1}}{2} \quad (1)$$

The results could only be reproduced using the 31-point scale, which was also used in the study by Arnold, et al. [1]. Table 2 shows the results of the Wilcoxon test. We chose to use a non-parametric test since not all scales met the requirements for a Welch test that was used in Arnold, et al. [1].

Table 2: Results of the priming. The priming effects were only reproduced using the 31-point scale.

	4-point	11-point	31-points	continuous
Wilcoxon test	W =140.5 p = .49	W = 185.5 p = .46	W = 229 p < .05	W = 143.5 p = .56

4. DISCUSSION

Our instructions aiming at varying the accuracy of ratings affected the subjects' behaviour in terms of the number of playbacks, as intended. In general, our subjects used a smaller number of playbacks than the subjects in the study by Jensen and Tøndering [4]. We conclude that a larger number of repetitions does not necessarily yield a greater rating accuracy.

Furthermore [4] reported that they received less extreme results when using scales with more levels. We observed that prominent syllables are equally likely to receive extreme rating values on all scales.

The correlations between the results obtained with different scales in the present study are not quite as high as reported by Jensen and Tøndering [4] but still strong. This difference is explicable since the prominence rating in their study was based on the word level. Prominence rating on the word level has been reported to be more robust than on the syllable level [6].

Subjects needed more time to complete their task when using a scale with more levels compared to the 4-point scale. The continuous scale required less time than the 11-point and 31-point scales.

The priming effect is apparently smaller in this study than in [1]. It is possible that the steps on the 4-point and 11-point scales are too small to replicate the priming effect observed in [1]. As for the continuous scale, subjects may have imposed different internal subdivisions, which may prevent the replication of the priming effect. In contrast, the 31-point scale evidently has the appropriate amount and spacing of steps to evoke the priming effect in line with that observed in [1].

5. CONCLUSION

We evaluated four scales for the rating of syllable prominence: 4-point, 11-point, 31-point scales, and a continuous scale. We found that subjects were able to use the full range of scales even with many levels. Using a scale with more levels enables good rating results and more interesting insights into inter-rater agreement. The priming effects reported by Arnold, et al. [1] could only be replicated using the 31-point scale.

6. REFERENCES

[1] Arnold, D., Wagner, P., Möbius, B. 2010. The effect of priming on the correlations between prominence ratings and acoustic features. *Speech Prosody 2010, Satellite Workshop on Prosodic Prominence* Chicago, IL.

[2] Eriksson, A., Thunberg, G., Traunmüller, H. 2001. Syllable prominence: A matter of vocal effort, phonetic distinctness and top-down processing. *Proceedings of Eurospeech 2001*, 399-402.

[3] Grover, C., Heuft, B., Coile, B.V. 1997. The reliability of labeling word prominence and prosodic boundary strength. *Proceedings of the ESCA Workshop on Intonation* Athens, 165-168.

[4] Jensen, C. Tøndering, J. 2005. Choosing a Scale for Measuring Perceived Prominence. *Proceedings of Interspeech 2005* Lisbon, 2385-2388.

[5] R Development Core Team 2010. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, *http://www.R-project.org/*

[6] Streefkerk B. 2002. *Prominence – Acoustical and Lexical/Syntactic Correlates*. Utrecht: LOT.

[7] Windmann, A., Wagner, P., Tamburini, F., Arnold, D., Oertel, C. 2010. Automatic prominence annotation of a german speech synthesis corpus: Towards prominence-based prosody generation for unit selection synthesis. *Proc. SSW7* Kyoto, 377-382.