Choosing a Scale for Measuring Perceived Prominence

Christian Jensen

Department of English
Copenhagen Business School, Denmark
cje.eng@cbs.dk

John Tøndering

Institute of Nordic Studies and Linguistics,
Department of Linguistics
University of Copenhagen, Denmark
johnt@hum.ku.dk

Abstract

Three different scales which have been used to measure perceived prominence are evaluated in a perceptual experiment. Average scores of raters using a multi-level (31-point) scale, a simple binary (2-point) scale and an intermediate 4-point scale are almost identical. The potentially finer gradation possible with the multi-level scale(s) is compensated for by having multiple listeners, which is a also a requirement for obtaining reliable data. In other words, a high number of levels is neither a sufficient nor a necessary requirement. Overall the best results were obtained using the 4-point scale, and there seems to be little justification for using a 31-point scale.

1. Introduction

The purpose of this paper is to evaluate the use of different scales for measuring the perceived prominence of syllables and words. In this investigation only word-level prominence is considered.

Words vary with respect to the degree to which they are felt to stand out from their surroundings. Some words are perceived as prominent, or emphasised, while others are perceived as less prominent.

Prominence, as perceived by groups of raters, has been measured on different types of scale: some use a 31-point scale from 0 to 30, first described in [1]. The strength of this scale is that it allows for very fine gradation of the perceived prominence, even for a single rater, but this also makes the task quite difficult – maybe too difficult for inexperienced raters. Others, e.g. [2, 3], have proposed to use instead a simple binary (2-point) scale (0 or 1) and use the cumulative (or average) score of each word as an expression of its level of prominence, which results in a much simpler task for the raters. The disadvantage of this simple scale is that it may force raters to conflate items which they perceive as "different, but within the same category". Potentially, this can lead to a reduced or lost ability to distinguish variations in perceived prominence at either end of the prominence continuum. For example, unemphatic, accented words and accented words with special emphasis may simply be judged as *prominent* by all raters, even if the words with special emphasis are generally felt to be more prominent. In addition, the level of gradation you achieve with this scale is directly proportional to the number of raters: to get the same gradation as is (potentially) possible with the scale from 0 to 30 you need 31 raters. As a possible compromise between these two scales one could use a 4-point scale (e.g. from 0 to 3) [4]. While this scale is much simpler than the 31-point scale it still allows raters to make some gradation in their prominence evaluations.

Perceived prominence is often associated with the linguistic categories accent and lexical stress. Proponents of a two-category system might therefore prefer the binary scale (accented or not), while a four-level scale might reflect a division into focused/nuclear accent, accent, secondary accent, and no accent. However, even assuming a hierarchical relationship between these elements this is different from our notion of prominence. Perceived prominence varies *within* these categories, and it seems that prominence, unlike the categories accent and lexical stress, is felt by listeners to be continuously variable, that is, a question of more or less. The variations in perceived prominence vary with linguistic meaning, especially semantic and pragmatic meaning such as information structure, and it is our contention that listeners respond to minor variations, even if a single listener cannot assign them consistently and accurately to discrete levels. These variations can instead be captured by using (mean) judgments from multiple listeners. Perceived prominence may also have to be considered in investigations of the acoustic correlates of stress and accent, or the automatic recognition of these categories.

We investigated the three prominence scales outlined above with the purpose of answering two overall questions: does the choice of scale influence the results with regard to 1) the perceived prominence relations of words in utterances, and 2) the ability to make observations about statistically significant differences between words. These questions were addressed from the point of view of three relevant linguistic parameters which are known to be associated with perceived prominence: *part of speech membership*, *information structure* and *correlation with* F_0.

2. Method

The speech material chosen to evaluate the scales was two short monologues from the Danish DanPASS project [5], both recordings of a map task activity. The two monologues, by two different male speakers, included a total of 123 words. The monologues were divided into 27 shorter phrases which were presented via a web page (one phrase per page). The raters could hear the phrase as many times as they wanted by pressing a "play" button, and indicated their judgment by clicking the appropriate scale point. Time consumption and a count of sound file playbacks were recorded for each phrase.

71 listeners participated in the experiment, most of them university students with little or no phonetic experience. They were randomly assigned to a specific scale. Seven raters had to be excluded because they are not native speakers of Danish, leaving 64 for further analysis: 24 on the 2-point scale, 21 on the 4-point scale and 19 on the 31-point scale. Since most tests required rater groups of equal size a random selection was made

of 19 raters from the 2- and 4-point scale groups for these comparisons. The instructions to the raters were presented from the web page and were identical for all three groups, except for the details about the specific scale. The concept of prominence was explained and exemplified, and raters were advised that prominence might be a question of "more or less". *0* represented *no prominence*, but no other scale points were defined. Prominent words could be assigned values *up to* the scale maximum. Raters using the 2-point scale were informed that they could not grade their ratings but were given a forced choice.

3. Results

3.1. Reliability

Note: the phrase "the 2/4/31-point scale" is used in the following as shorthand expressions of "the prominence ratings obtained from the group of listeners using the 2/4/31-point scale".

The reliability of the data was tested by calculating Cronbach's α coefficient, and the results are displayed in Table 1.

Scale	*Cronbach's* α
2-point	0.961
4-point	0.961
31-point	0.940

Table 1: *Reliability coefficients*

The coefficients, which express the extent to which the scores of the individual raters covary, are high for all three groups, or scale types, and the difference between them is non-significant (M = 1.02, $p > 0.05$).

3.2. Comparison of prominence ratings

The first question to be addressed is whether the prominence ratings on the three scales express the same relations between words. In order to be able to make direct comparisons all scores were normalised by dividing each value with the scale maximum (1, 3 or 30, respectively), which fits all data to a normalised scale of 0 to 1 without affecting the relations between scores. These values were then plotted on a line chart for simple visual inspection. An example diagram of one phrase is shown in Fig. 1.

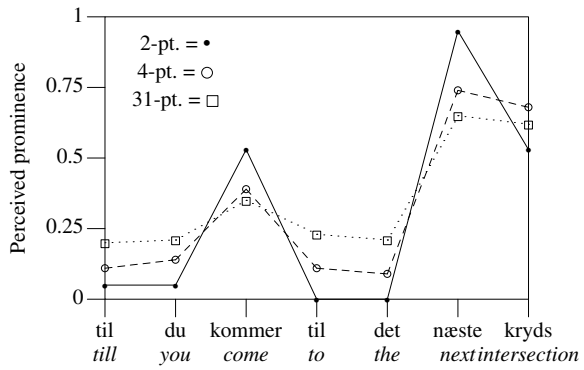

Figure 1: *Prominence of selected phrase – all scales*

The diagrams showed a high level of agreement across the three scales, which was further tested in a correlation analysis (Spearman's ρ). The result can be seen in Table 2.

Correlation	*4-pt*	*31-pt*
2-pt	0.933	0.926
4-pt	—	0.964

Table 2: *Correlation coefficients (Spearman's ρ) across all three scales*

The correlation coefficients were high for each scale pair and quite similar, with the best correlation apparently between the 4-point scale and the 31-point scale. The preliminary conclusion is clear: raters arrive at approximately the same rank order of perceived prominence regardless of the scale used.

It appears from Fig. 1 that the 2-point scale displays somewhat larger variation in values between the scale minimum and maximum than the 4-point scale and especially the 31-point scale. This was in fact a general trend demonstrating a certain compression of values on the 31-point scale (and to a lesser degree the 4-point scale), while the 2-point scale has more mean values near the scale extremes. Analyses of the distribution of scores (inter-quartile range for each rater and visual inspection of *x*-*y* plots) showed that many raters on the 31-point scale assigned most ratings to a restricted – sometimes very restricted – range of the scale, either at the lower, the middle or the higher end of the scale. There are therefore no *mean values* at the scale extremes, although there were many individual scores near the minimum and maximum values.

As an attempt to compensate for the overall difference in distribution of mean values a further transformation of the data was performed: the lowest recorded mean value was set to 0 and other values were scaled linearly so that the highest mean value was 1. While this transformation did smooth out some of the differences between the scales they did not disappear entirely. All further analyses are therefore performed on raw data.

3.3. Obtaining statistically significant differences

One very important aspect of choosing a scale is whether it will affect the ability to obtain statistically significant differences between test items. The hypothesis might be that scales with too few points (most notably the 2-point scale) would mask subtle perceptual differences which could be brought out with more scale points.

This suitability of the three scales for quantitative analysis was tested by examining the association between perceived prominence and three linguistic phenomena: part of speech membership, information structure and a specific acoustic correlate, namely F_0. The purpose was to see if the data obtained by using three different scales will lead to different conclusions about linguistic structure.

3.3.1. Comment on the statistical procedures

It is not possible to make direct comparisons of prominence ratings across different scale types, as different statistical procedures are required for the three scales. What we have done instead is to use the statistical method which is found appropriate for the specific scale and examine what the combination of (responses on) a specific scale and the associated statistical method will allow us to say about the prominence relations in the utterances. This resembles quite well the situation in which researchers find themselves when they are making a choice about scale type.

For all scales we have decided to use nonparametric methods. There is a great deal of disagreement in the literature about

	Scale →		2-point		4-point		31-point	
	Part of speech	*n*	*Ranked*	$\bar{x}$	*Ranked*	$\bar{x}$	*Ranked*	$\bar{x}$
1	Adjectives	9	*adj*	0.92	*adj*	0.73	*adj*	0.67
2	Nouns	28	*n*	0.78	*n*	0.66	*n*	0.63
3	Interjections	3	*int*	0.60	*int*	0.50	*int*	0.58
4	Adverbs	12	*adv*	0.58	*adv*	0.38	*adv*	0.40
5	Verbs	13	*v*	0.34	*v*	0.30	*pron*	0.35
6	Pronouns	16	*pron*	0.33	*pron*	0.30	*v*	0.35
7	Conjunctions	10	*conj*	0.17	*prep*	0.21	*prep*	0.28
8	Articles	2	*art*	0.13	*conj*	0.13	*conj*	0.24
9	Prepositions	30	*prep*	0.10	*art*	0.12	*art*	0.22

Table 3: Prominence ratings and parts of speech. Left braces indicate *non*significant differences. Non-adjacent, nonsignificant differences on the 31-pt scale: adv-v, art-prep

whether scales like the ones in this investigation should be considered continuous or ordinal scales, but following [6] we have decided to go with the perhaps more traditional or conservative choice of considering them ordinal scales, and so use nonparametric methods. For significance testing on the 2-point scale we use the Fisher exact test or a chi-square test with corrections for continuity (when $n > 40$), and for the other two scales we use the Wilcoxon-Mann-Whitney test with correction for ties (WMW).

3.3.2. *Parts of speech*

The mean prominence ratings of nine parts of speech are listed in Table 3, ordered according to their ranking on each scale. These ranking are very similar for all three scales. The only difference which can be detected is the relegation of *prepositions* to ninth place on the 2-point scale, instead of the seventh place it holds on the other two scales. (The different ranking of *pronouns* and *verbs* on the 31-point scale is irrelevant.) Most of the differences between the classes are significant: except for two cases on the 31-point scale (see the table caption) all differences between classes which are not adjacent in the rankings are significant, and of the differences between adjacent classes four are nonsignificant on the 2-point scale, two are nonsignificant on the 4-point scale, and three are nonsignificant on the 31-point scale (giving a total of five differences which are not significant for this scale). These figures are quite similar, with a small bias in favour of the 4-point scale, where the highest number of significant differences was found.

3.3.3. *Information structure*

According to many theories of information structure new information is either expected, or even specified, to be the most prominent word in a phrase. One of these theories is [7] which was applied to the data in this study, with the modification that there may be more than one new idea expressed in a phrase. 15 of the 27 phrases in the study contain new information, five of them contain two separate ideas, and for each of the 20 resulting pairs the rating of the most prominent word signalling new information was compared to the rating of the most prominent word carrying non-new information (given or accessible in Chafe's terms), thus testing the hypothesis that new information is more prominent than other information (H_1). H_0 states that the perceived prominence of the new information is less than or equal to that of the given/accessible information. The result of the comparisons is displayed in Table 4.

In four cases (three on the 31-point scale) the new information is not more prominent than the non-new information, in which case H_0 cannot be dismissed. Of the remaining 16 (17) cases, where the new information had higher prominence ratings than the non-new information, nine were significant on the 2-point scale (Fisher exact test, one-tailed, $p < 0.05$); 15 were significant on the 4-point scale and 14 on the 31-point scale (WMW, one-tailed, $p < 0.05$).

	2-pt	*4-pt*	*31-pt*
n	16	16	17
new > not new	9	15	14

Table 4: *Significant differences between prominence ratings of words carrying new versus non-new information*

Here we find a clear difference between the 2-point scale and the 4-point and 31-point scales in the number of significant differences. Our conclusion about the relative prominence levels of new versus non-new information would therefore be affected by our choice of scale, provided that we want to verify observed differences in mean ratings statistically.

3.3.4. *Correlation with F_0*

The prominence level of a Danish accented syllable, and of the word in which it occurs, is generally felt to be associated with, among other cues, a rise in F_0. The greater the rise, the more prominent the syllable is perceived to be. For this investigation two F_0 values were measured for all words in which such a rise occurs: the F_0 trough and the F_0 peak value within the domain of onset of the accented vowel and the end of the word (since we were concerned with word level prominence). The rise is expressed as the difference in semitones between these two values, and the values for the rises were then correlated against the prominence ratings from the three scales. The results are displayed in Table 5.

Scale	ρ
2-pt	0.593
4-pt	0.626
31-pt	0.606

Table 5: *Correlation (Spearman's ρ) between perceived prominence and F_0*

The correlation coefficients are very similar for the three data sets, indicating the the association between prominence and F_0 can be described equally well regardless of the scale used. To the (slight) extent that any difference can be detected

it seems that the correlation is better with data obtained on the 4-point scale.

3.4. Rater effort, or level of difficulty

In a few places we have described the 2-point scale, and to some extent the 4-point scale, as "simpler" and less difficult for the rater than the 31-point scale. At least this was our expectation, and as an attempt to capture this we measured the time consumption for each phrase and the number of times the raters listened to each phrase. The hypothesis is that both of these measures will increase with an increase in the number of scale points. The results can be seen in Table 6.

Scale →		2-point	4-point	31-point
Time consumption (sec)	*Mean/ phrase*	24.1	28.4	34.3
	Mean/ word	5.3	6.2	7.5
	Index	100	118	142
Number of playbacks	*Mean/ phrase*	2.5	2.6	2.9
	Index	100	104	115

Table 6: Time consumption and number of sound file playbacks

As predicted, there is a difference in the average time raters spent on rating a phrase across the three scales. This might be expressed as an increase of 18% when going from two to four scale points, and an increase of 42% when going from two to 31 points. All pairwise comparisons between the three scales are significant (t-tests, one-tailed, $p < 0.05$). The pattern is less clear for the number of playbacks, where only the tendency for more playbacks on the 31-point scale compared with the 2- and 4-point scales is statistically significant.

It must be concluded, though, that using more scale points will result in a somewhat higher "cost".

4. Discussion and conclusion

Two main questions were asked about the influence of scale type on ratings of perceived prominence: 1) do we get the same prominence relations in utterances, as expressed in mean values and rankings, and 2) does scale type affect our ability to make observations about statistically significant differences between words. The overall conclusion must be that the perceived prominence relations in the utterances are very similar whether expressed on a 2-point scale, a 4-point scale or a 31-point scale. The differences are small and are mostly caused by a tendency for some raters to prefer a restricted range within a multi-level scale. The differences are also relatively small when it comes to statistical testing of observations, but it does seem that raising the number of scale points from two to four yields slightly better results: there are more significant differences between the part of speech categories and between words with new versus given/accessible information, and the correlation with F_0 is better. No such improvement can be obtained, however, by raising the number of scale point to 31. On the contrary we find slightly fewer significant differences on this scale.

One reason for this finding may be that it is too difficult for untrained listeners to use the 31-point scale. In a parallel experiment (to be reported elsewhere) we had five expert listeners rate the same phrases as in this experiment (with slightly different instructions). The performance of this group was generally better than any random group of five untrained listeners (higher Cronbach α coefficient and more significant differences), which indicates that they did in fact do better on this scale. The analysis also showed, however, that five expert listeners cannot replace a larger group of untrained listeners if the objective is to find statistically significant differences – the number of observations becomes too small.

It was shown that "expenses", in terms of especially time consumption, grew with an increase in the number of scale points. Combined with the above observations this points to a recommendation of using many listeners rating on a scale with relatively few levels. A 2-point scale may then be adequate for most purposes and makes for the simplest and fastest task, but it would appear that increasing the number of levels to four results in slightly better performance. There seems to be no justification for using a 31-point scale, unless the requirement of using many listeners cannot be met. The task becomes more difficult and takes more time, and there is no gain in terms of precision or "discriminatory power" to balance the extra cost.

5. References

[1] Fant, G. and Kruckenberg, A., "Preliminaries to the study of Swedish prose reading and reading style", STL-QPSR 2/1989:1–83, 1989.

[2] Wightman, C., "Perception of multiple levels of prominence in spontaneous speech", ASA 126th Meeting Denver 1993 (abstract).

[3] Streefkerk, B. M., Pols, L. C. W. and ten Bosch, L. F. M., "Towards finding optimal features of perceived prominence", Proc. 14th International Congress of Phonetic Sciences, pp. 1769-1772, San Francisco, 1999.

[4] Jensen, C., "Stress and Accent. Prominence relations in Southern Standard British English", PhD thesis, University of Copenhagen, 2004.

[5] Grønnum, N., "DanPASS – Danish Phonetically Annotated Speech", Apr. 2005; http://www.cphling.dk/pers/ng/danpass.htm.

[6] Siegel, S. and Castellan, N. J., "Nonparametric statistics for the behavioral sciences", New York: McGraw-Hill, 1988.

[7] Chafe, W., "Discourse consciousness, and time: the flow and displacement of conscious experience in speaking and writing", Chicago: The University of Chicago Press, 1994.