

# Perception of prominence patterns in English nominal compounds

Gero Kunter

Department of English, Universität Siegen, Germany  
kunter@anglistik.uni-siegen.de

## Abstract

This paper investigates prominence patterns in English noun-noun compounds. A perception experiment is presented in which naive listeners rated the prominence relation between the two elements. It is found that either the first or the second element of the stimuli is perceived as more prominent, and that the distinction can be considered categorical. Pertinent statistics are discussed that can assess the reliability of different raters. It is shown that right prominence is more difficult to perceive by many listeners than left prominence.

**Index Terms:** English compounds, prominence patterns, perception

## 1. Introduction

In English, compounding is a highly productive process to create new words such as *boarding schools*, *eviction notice* or *state colleges*. While most noun-noun compounds are considered to have highest prominence on the left element, some constructions have a prominence pattern that is similar to that of syntactic phrases [1], in which the right element is considered to be more prominent. Example (1) lists instances of the first type; the compounds in (2) are examples of the second:

- (1) *greenhouse, textbook, word-processor*
- (2) *paper cup, summer camp, Boston marathon*

While the factors that determine the distribution of these two patterns have been variously discussed (e.g. [1, 2, 3]), only little attention has been given to the question whether this description of prominence patterns in English compounds is indeed appropriate from a perceptual point of view. [4] is one of the few empirical studies available. They found that listeners seem to be able to correctly identify whether speakers intend to produce a compound of the first or of the second type. However, a forced-choice paradigm was applied in their study, so it remains unclear whether we are dealing with a categorical or gradient phenomenon, and if it is categorical, how many classes one should assume. Furthermore, with only three pairs of constructions with contrasting prominence patterns, the number of stimuli was too small to generalize on prominence patterns in English compounds. Finally, [4] report a considerable degree of variation between the six participating raters, whose nature is not explored.

The present paper presents a perception experiment in which naive listeners attempt to rate the prominence relation between the two elements in 105 English noun-noun compounds. The results show that, first, listeners are not equally proficient in assessing the prominence patterns of compounds. Suitable statistical methods will be presented that can estimate rater proficiency. Second, English compounds are indeed either left-prominent or right-prominent, and the distinction between the

two patterns should be considered categorical. Third, the perception of right prominence is more difficult than the perception of left prominence.

## 2. Data

31 undergraduate students of University of California, Santa Cruz (all native speakers of standard American English) participated in the perception experiment, none of whom had a special training in speech phonetics or linguistics. This use of naive test subjects has the advantage that no prior knowledge about the acoustic correlates of prominence as related to stress could interfere with the ratings. Other studies involving prominence judgment tasks (e.g. [5, 6]) have reported that trained and naive subjects may provide significantly different ratings. [6] conclude that the difference is due to a “non-linguistic awareness of the phonetic properties”, and thus the trained subjects “responded to the stimuli in a less natural way than the [...] naive subjects”. Such a less natural response is clearly undesired in an experiment such as the present.

A speech corpus, consisting of high-quality studio recordings from seven different radio news-speakers [7], provided the 105 different noun-noun compounds that constituted the stimulus set (15 compounds per speaker). The stimuli were presented both acoustically and as text on a computer. Participants were instructed to judge the perceived prominence of the compound elements by placing a graphical marker on an unmarked scale that was presented on the computer screen. The center of the scale was introduced to the participants as representing a stimulus in which both elements were of equal prominence, while a position on the left-hand side or right-hand side of the scale was defined to represent a more prominent left or right element, respectively. Participants were encouraged to take full advantage of the available space: if, for instance, the first element was clearly more prominent than the second element, participants were expected to place a marker on the far left side of the available space. The marker positions were later mapped onto integer numbers ranging from  $-499$  to  $500$ . Hence, negative values corresponded to a higher perceived prominence of the left element, and positive ratings to a higher perceived prominence of the right element.

Raters differed considerably in the scale range that they used for their ratings. For instance, ratings of one participant covered 910 units on the internal scale (from  $-452$  to  $458$ ), while the ratings of another participant spanned between  $-255$  and  $272$ . Such individual differences are well-known to occur in rating experiments using unmarked scales [8], and are generally held to reflect differences in rating strategies rather than relevant perceptual differences. To counter these effects, a quantile transformation [9] was applied in which all ratings for a given speaker was divided by the standard deviation of ratings for that speaker. The interpretation of negative and positive values as

judgments of left and right prominence, respectively, is retained in this transformation.

### 3. Rater reliability

It is generally assumed that any given English compound is either left-prominent or right-prominent, and that a native listener should be capable of identifying this prominence relation. In this case, all raters in the perception experiment should show a very high degree of rating consistency. However, it may also be the case that there are individual proficiency differences: some listeners might be more capable of identifying prominence patterns than others. It is also possible that there are different perceptual strategies when assessing compound prominence patterns. For instance, a given compound may be considered as right-prominent by one group of raters, but as left-prominent by a different group. In this case, a rater would show a high degree of consistency if compared to other raters that apply the same strategy, but would not be consistent with ratings provided by raters from a group with a different strategy.

The cluster analysis in this section uses the intraclass correlation coefficient (*ICC*, [10]) to assess the degree of rater proficiency, and to determine whether evidence for different rating strategies can be found.

#### 3.1. Procedure

The *ICC* expresses the degree of group-internal consistency, i.e. consistency of multiple responses across different stimuli. As with other correlation coefficients, a value of 0.0 represents no within-group consistency whatsoever, and a value of 1.0 corresponds to full consistency between all group members. In the present analysis, it is used as a similarity metric in an agglomerating hierarchical cluster analysis. The analysis first calculates the *ICC* for all pairwise combinations of raters, and detects the pair with the highest *ICC*. This pair is joined into a cluster. In all subsequent steps, *ICCs* are calculated for all pairwise combinations of remaining raters and the averaged rating within each existing cluster, and again, the combination with the highest *ICC* is joined into a new cluster. This is repeated until all raters are accounted for.

If there are no considerable differences in rating proficiency across individuals, the resulting tree should show a very flat structure, with all raters added to the structure at very high *ICC* levels. If there are individual proficiency differences, the addition of some raters should considerably reduce the *ICC* at that stage. Finally, if speakers resort to different rating strategies, we expect to find a tree structure with two or more clusters with high internal consistency that are joined only late in the analysis.

#### 3.2. Results

Figure 1 shows the tree structure that emerges from the cluster analysis. The *ICC* level at which the clusters were joined is given on the y axis. Raters 8 and 12 on the left-hand side have the highest degree of consistency of all pairs ( $ICC = 0.840$ ). As the analysis proceeds in adding further raters to the structure, consistency decreases only slightly to  $ICC = 0.697$  after 17 raters are joined in this cluster (indicated by the lower dotted line). Apparently, there is a strong agreement of ratings for any given compound within this group.

Addition of further raters to this cluster introduces a notable decrease in consistency. Starting with rater 24, *ICC* levels quickly deteriorate during the next six additions (from 0.639

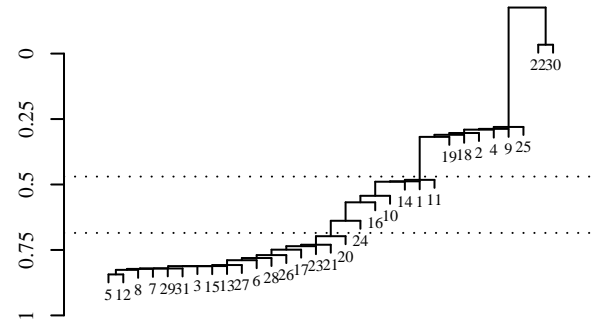


Figure 1: *Agglomerating cluster analysis using ICC as similarity metric (N = 31).*

to 0.482 for rater 11). Thus, these six raters (shown between by the two dotted lines) provide ratings that agree only to a medium degree with the large group of 17 consistent raters. The remaining participants provide quite inconsistent ratings: *ICC* levels drop from 0.318 to 0.280 between rater 19 and 25, which suggests that the rating behaviors of these participants disagree very strongly from those of any other rater. Finally, responses from raters 22 and 30 deviate so far from those of all other raters that for them, the *ICC* is insignificant.

In sum, there is a large group of 17 raters with strongly agreeing responses, resulting in very high *ICC* levels in this cluster. Another group consisting of six raters has only a medium agreement with the averaged data from the larger group, while ratings from the remaining eight participants show only little or no consistency with the averaged ratings from all other raters.

The emerging tree structure does not suggest that there are different rating strategies: there is a single initial cluster with a very high *ICC* to which all other raters are added in the subsequent steps of the analysis. A second rating strategy would need to form another consistent group, yet there is no other cluster with a comparable high degree of internal consistency,

#### 3.3. Discussion

Using intraclass correlations to determine the degree of rating agreement between different raters, the present cluster analysis has found a large group of 17 participants who provide highly consistent ratings when asked to judge the prominence relation in English compounds. The remaining participants deviate considerably from the ratings provided by the members of this group.

The linear tree structure, in which raters are generally added subsequently to the initial cluster, implies that these rating inconsistencies are not due to different perception strategies, otherwise, the analysis would have found a second cluster with high internal consistency apart from the one formed by the 17 raters. Apparently, rating inconsistencies are rather due to individual, random variation than to different rating patterns.

Thus, the analysis of rating consistency implies that only slightly more than half of the participants may be considered as proficient judges of left- and right-prominence in the compound stimuli whose data can be assumed to reflect the prominence relation reliably. With decreasing rating consistencies, prominence judgments from the remaining participants are found to

be less and less reliable, up to the point where there is only accidental agreement in the ratings.

Hence, contrary to the stance found in much of the pertaining literature (e.g. in [4]), it is a non-trivial task to judge the perceptual prominence in English compounds, and similar individual proficiency differences are to be expected in other experiments that investigate the prominence structure within spoken data. Such experiments may profit from statistical methods like the one discussed here, as they can be used to filter out the potentially noisy data from less proficient raters, and to identify those participants that provide reliable prominence ratings.

## 4. Prominence patterns in compounds

To address the question raised in the introduction whether perceptual prominence in English compounds should be considered a categorical or a gradient phenomenon, this section investigates the distribution of prominence ratings as provided by the 17 reliable participants in the perception experiment.

### 4.1. Results

For each compound, the median of the prominence ratings was calculated. The resulting distribution of median prominence ratings is displayed in Figure 4.1 in the form of a density plot. The solid line corresponds to the distribution obtained from the proficient raters, while the distribution of ratings from all participants is shown for illustration by a dashed line. The height of the two lines corresponds to the relative frequency of ratings on the respective region of the perceived prominence rating scale.

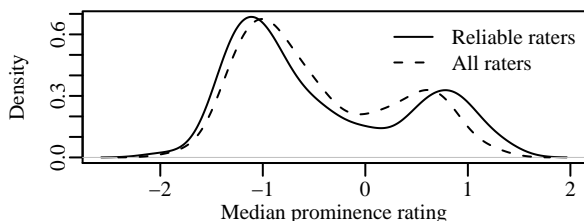


Figure 2: Distribution of median prominence ratings.

The solid density plot shows two clear peaks, one on the negative side of the scale (corresponding to stimuli where the left element was perceived as more prominent) and another peak on the positive hand side (indicating prominence of the right element). Examples of stimuli with ratings indicating left prominence *water pipes* (-1.145) and *correction officials* (-1.058), while the ratings for *taxation committee* (0.959) and *senate president* (0.980) indicate right prominence. The two peaks are separated by a valley at the scale center, which suggests that only very few stimuli are perceived as having equal or similar prominence of both elements.

A hierarchical cluster analysis of the median prominence ratings (using Ward clustering with Euclidean distance measure) yields only two clusters that are separated at a median prominence rating of 0.25. The larger cluster (76 stimuli) has an average prominence rating of  $-0.910$  ( $SD = 0.442$ ), which corresponds to a clear perceptual prominence of the left element. For the smaller cluster (29 stimuli), the average is  $0.814$  ( $SD = 0.263$ ), corresponding to higher perceptual prominence of the right element. Neither the density plot nor the cluster analysis provide evidence for a third prominence pattern.

## 4.2. Discussion

The distribution of prominence ratings supports the two prominence categories assumed for English noun-noun compounds: these constructions may either be left-prominent or right-prominent. In agreement with estimates in the literature [11], left prominence is by far the more frequent pattern, but right prominence occurs in about 30 percent of data. The density plot and the cluster analysis imply that there is no other prominence pattern apart from these two; in particular, there is no indication that there is a group of compounds in which both elements have equal prominence. The few stimuli with median prominence ratings close to zero seem to be within the expected range if we assume that ratings for stimuli belonging to either pattern are normally distributed.

As there is only little overlap between the two perceptual patterns, evident in the very deep valley between the two peaks, it seems that the proficient raters perceive the two prominence patterns in a categorical way, and that assignment of a stimulus to either of the two available patterns is a straightforward decision for these raters. However, the density plot for all 31 participants (indicated by the dashed line) suggests that this is not true for all raters. Here, not only is the right peak closer to the scale center, the separating valley is also much less pronounced, which suggests more overlap between the two categories. Yet, on the left-hand side of the scale, the two distributions are nearly identical, so it seems that right-prominent stimuli are particularly difficult to judge by less proficient raters. This hypothesis is discussed in the next section.

## 5. Prominence patterns and proficiency

If the prominence relation in right-prominent compounds is indeed more difficult to perceive for the less reliable raters, as suggested by the comparison of the two density plots in Figure 4.1, it seems plausible to predict a higher degree of rating variability for stimuli that have a higher median prominence rating. For proficient participants, it is expected that rating variability remains mostly constant for all median prominence ratings.

A non-parametric regression [12] with the median prominence ratings as predictor and the interquartile range for each stimulus (*IQR*) as the response variable is used to test these predictions. The *IQRs* represent the distribution of ratings around the median, and thus are a measure of rating variability for each stimulus.

### 5.1. Results

The scatter-plots in Figure 5.1 show the relation between median prominence ratings (on the horizontal axis) and *IQR* (on the vertical axis). The left panel refers to the ratings obtained from all participants, while the right panel is restricted to the proficient raters. The line in each panel is the non-parametric regression line that estimates the influence of median prominence ratings on the *IQRs*, while the shaded area is the reference band for a ‘no effect’ model [12]. If the regression line runs within this band, there is no reason to assume a relation between the two variables; if the deviation from this area is sufficiently large, there is a significant effect of median prominence ratings on the *IQRs*.

Apparently, the regression line on the right panel is lower than that on the left panel, which agrees with the finding above that their overall amount of variability decreases if we consider only the proficient raters. In addition to this general observation, we find a significant deviation of the regression line

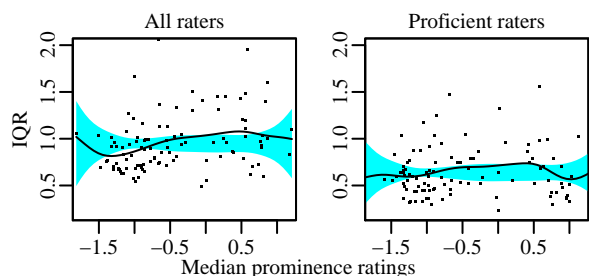


Figure 3: Scatter-plots of median prominence ratings by IQR.

from the reference band in the left panel (test for ‘no effect’ model:  $p = 0.043$ ). The *IQR* is slightly lower than expected for left-prominent stimuli, and hence the regression line is below the reference band. With increased prominence ratings, the *IQR* increases as well, until it is significantly higher for right-prominent stimuli. There is no evidence for a comparable effect in the data obtained from the proficient rater group in the right panel (test of ‘no effect’ model:  $p = 0.176$ ).

## 5.2. Discussion

The non-parametric regression reveals that there is particularly strong variability in the prominence ratings among the whole group of participants if the compound is generally perceived as right-prominent, while for left-prominent compounds, rating variability reaches a level that is comparable to that shown by the proficient raters across the whole range of perceived prominence relations. Proficient judges in contrast perform similarly well for both patterns.

One explanation for this result may lie in the underlying perceptual mechanism. It has been argued [13] that right prominence is phonologically realized by placing a pitch accent on both elements. These compounds are generally perceived as right-prominent due to a psycho-acoustic compensation of pitch declination [5]. Left-prominent compounds are held to have a pitch accent only on the left element. Recent acoustic studies of prominence patterns in compounds, e.g. [14], provide strong support for this assumption. Thus, left prominence may be fairly easy to perceive and classify by listeners, because the classification involves a comparison between accented and unaccented material, i.e. between linguistic items that belong to two different phonological categories. In contrast, right prominence may be more challenging to perceive correctly because it involves the comparison of two elements that belong to the same phonological category. Studies of other perceptual categories [15, 16] have shown that discrimination accuracies usually decrease considerably in within-category comparisons if the distinction is perceived categorically. The present findings suggest that this may also be the case for accent perception in English compounds.

## 6. Conclusions

This paper discussed statistical methods which assess the proficiency of participants in a perception experiment. Using data from only the proficient raters, it was shown that in English compounds either the left or the right element is perceived as more prominent, and that the distinction between these two patterns seems to be a categorical one. The analysis suggested that the less proficient raters have particular difficulties in providing reliable prominence ratings in the case of right-prominent

compounds, a finding that may be explained by the fact that both elements are accented in these stimuli. Thus, the different methods shown in this paper may offer insights into between-listener differences in auditory perception, which may be used to increase the reliability of empirical studies.

## 7. Acknowledgments

This work was supported by grants PL151/5-1 and PL151/5-3 from the *Deutsche Forschungsgemeinschaft* (DFG). Thanks are due to Ingo Plag for fruitful discussions and comments, and to the three anonymous reviewers for their very valuable input that helped to improve this paper.

## 8. References

- [1] M. Y. Liberman and R. Sproat, “The stress and structure of modified noun phrases in English,” in *Lexical matters*, I. A. Sag and A. Szabolcsi, Eds. Stanford: Center for the Study of Language and Information, 1992, pp. 131–181.
- [2] H. J. Giegerich, “Compound or phrase? English noun-plus-noun constructions and the stress criterion,” *English Language and Linguistics*, vol. 8, pp. 1–24, 2004.
- [3] I. Plag, G. Kunter, M. Braun, and S. Lappe, “The role of semantics, argument structure, and lexicalization in compound stress assignment in English,” *Language*, vol. 84, no. 4, 2008.
- [4] E. Farnetani, C. T. Torsello, and P. Cosi, “English compound versus non-compound noun phrases in discourse: an acoustic and perceptual study,” *Language and Speech*, vol. 31, no. 2, pp. 157–180, 1988.
- [5] C. Gussenhoven and A. C. M. Rietveld, “Fundamental frequency declination in Dutch: testing three hypotheses,” *Journal of Phonetics*, vol. 16, pp. 355–369, 1988.
- [6] D. R. Ladd, J. Verhoeven, and K. Jacobs, “Influence of adjacent pitch accents on each other’s perceived prominence: two contradictory effects,” *Journal of Phonetics*, vol. 22, pp. 87–99, 1994.
- [7] M. Ostendorf, P. Price, and S. Shattuck-Hufnagel, *Boston University Radio Speech Corpus*. Philadelphia: Linguistic Data Consortium, 1996.
- [8] J. B. Taylor and H. A. Parker, “Graphic ratings and attitude measurement. a comparison of research tactics,” *Journal of Applied Psychology*, vol. 48, no. 1, pp. 37–42, 1964.
- [9] K. Jajuga and M. Walesiak, “Standardisation of data sets under different measurement scales,” in *Classification and information processing at the turn of the millennium*, R. Decker and W. Gaul, Eds. Berlin: Springer, 2000, pp. 105–112.
- [10] P. E. Shrout and J. L. Fleiss, “Intraclass correlations: uses in assessing rater reliability,” *Psychological Bulletin*, vol. 86, no. 2, pp. 420–428, 1979.
- [11] R. Sproat, “English noun-phrase accent prediction for text-to-speech,” *Computer Speech and Language*, vol. 8, pp. 79–94, 1994.
- [12] A. W. Bowman and A. Azzalini, *Applied smoothing techniques for data analysis. The kernel approach with S-Plus illustrations*. Oxford: Clarendon Press, 1997.
- [13] C. Gussenhoven, *The phonology of tone and intonation*. Cambridge: Cambridge University Press, 2004.
- [14] G. Kunter, “Compound stress in English. the phonetics and phonology of prosodic prominence,” Ph.D. dissertation, Universität Siegen, 2009.
- [15] A. M. Liberman, K. S. Harris, H. S. Hoffman, and B. C. Griffith, “The discrimination of speech sounds within and across phoneme boundaries,” *Journal of Experimental Psychology*, vol. 54, no. 5, pp. 358–368, 1957.
- [16] B. H. Repp, “Categorical perception: issues, methods, findings,” in *Speech and language. Advances in basic research and practice*, N. J. Lass, Ed. Orlando, FL: Academic Press, 1984, vol. 10, pp. 244–335.