Obtaining prominence judgments from naïve listeners – Influence of rating scales, linguistic levels and normalisation

Denis Arnold [1], *Petra Wagner* [2], *Bernd Möbius* [3]

[1] Quantitative Linguistics, University of Tübingen, Germany
[2] Faculty of Linguistics and Literature, Bielefeld University, Germany
[3] Department of Computational Linguistics and Phonetics, Saarland University, Germany
denis.arnold@uni-tuebingen.de, petra.wagner@uni-bielefeld.de, moebius@coli.uni-saarland.de

Abstract

In this paper we examine different approaches to obtain judgments of perceptual prominence. We discuss different prominence scales, the influence of the linguistic unit on which prominence is rated and the normalisation of prominence judgments. We propose the use of a multilevel scale for obtaining prominence judgments. It seems that naïve listeners can rate word prominence better than syllable prominence, resulting in better correlations to acoustics. It is shown that normalisation should be applied to the obtained ratings.

Index Terms: prosody, prominence, methods, normalisation, acoustic correlates

1. Introduction

It is widely accepted that prominence is a perceptual construct that describes the perceived strength of a given linguistic unit in the context of its neighbours. Questions remain whether prominence is gradual or categorical, whether prominence should be labeled on the word or syllable level, and also whether there is something like "absolute" prominence or if prominence has to be seen as relative to its context. Unfortunately, many papers dealing with prominence do not offer, or refer to, a definition of prominence. As pointed out by [1], many results in studies on prominence might be influenced by the approach chosen by the respective authors. In our studies, we refer to the definition given in [2]. Here, prominence is defined as the gradual perceived strength of a linguistic unit relativ to its environment.

A variety of methodological approaches have been used in prominence research. Experiments on natural [3, 4] and manipulated stimuli [5, 6] were conducted. Production experiments were carried out [1] as well as work using corpora with annotated prominence [7, 8]. Some studies dealt with the prediction of prominence [9], and some focused on the automatic annotation of prominence from the signal [10, 11] and others on the systematic differences between automatic and human annotated prominence [8]. There are also studies examining the influence of linguistic knowledge on the perceptual prominence [4, 12, 13]. This set of studies used a range of different scales, measure prominence on different linguistic levels and also have different conceptualisations of prominence. Moreover, the research has targeted different languages. While different approaches, like experimental vs. corpus work, are valuable to examine different questions, one has to be very careful when comparing the results of studies that use different methods to capture prominence.

In section 2 and 3 we will summarise the results of [17, 18]. In [17] 216 subjects rated 15 sentences in German with four different rating scales. In [18] 36 subjects rated word prominence versus syllable prominence on 15 sentences in German. In section 4 we will report results based on a normalisation of the data from these studies.

2. Evaluation of prominence rating scales

Many different rating scales have been employed in prominence research. A large number of studies use a binary scale, e.g. [14, 15, 16]. The clear advantage of this procedure is that it is easy to use for the annotators. It has been argued by [14] that with *n* raters one gets a n-level scale of prominence. However, this leads to yet another problem. If one uses the number of raters saying a given unit is prominent, one confuses the amount of prominence with the confidence of the rating. With a multilevel scale one can see how much prominence is assigned to a given unit and how confident the annotators are. Our data [17, 18] indicate that high confidence is not equal to extreme prominence ratings and vice versa. A variety of multilevel scales were used in the literature including a 3-point [8], 4-point [19], 11-point [20] and 31-point scale [3, 12, 13]. Two studies that focused on the use of scales found contradicting results [21, 19]. Grover et al. found that scales with more levels result in more reliable results [21], while Jensen and Tondering prefer a 4-point scale [19]. The results obtained with the three tested scales – binary, 4-point and 31-point scale - do

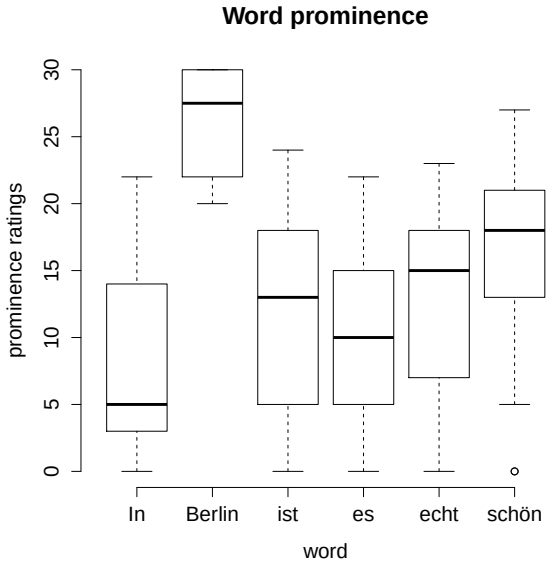

Figure 1: *Word prominence rating of a sentence in study [18]. In Berlin ist es echt schön - It is really nice in Berlin.*

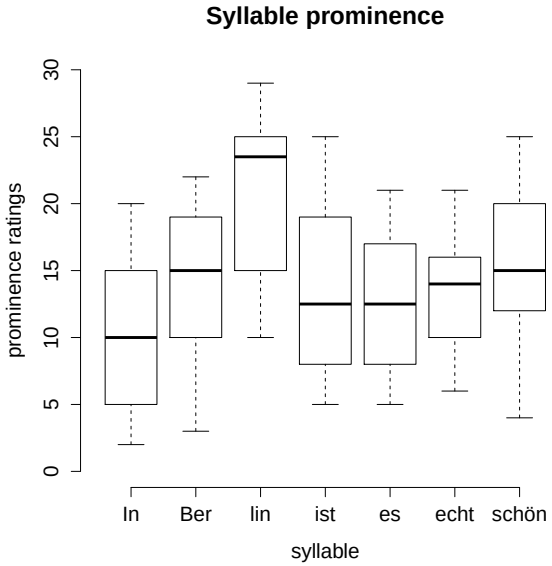

Figure 2: *Syllable prominence of a sentence in study [18]. In Berlin ist es echt schön - It is really nice in Berlin.*

not differ much. However, the authors argue that the 31-point scale is harder to use for naïve listeners, who cannot fully exploit it. For this reason, one finds less pronounced prominence patterns with the 31-point scale. In [17] we presented evidence that the use of multilevel scales, like a 11- or 31-point scale, has advantages over the use of a 4-point and a continuous scale for rating prominence. No evidence was found for the arguments that multilevel scales cannot be fully utilized by naïve listeners, that an increasing number of levels on a scale results in less pronounced prominence profiles, or that an larger number of levels is harder to use.

3. Prominence on different linguistic levels

The different aspects of perceptual prominence have been examined in studies on the word and syllable level. A systematic variation of the linguistic unit was only reported in [14]. In [18] we presented results suggesting that there is no simple relation between the prominence rated on the word level and syllable prominence ratings. We found that complex interactions influence the assigned prominence on the different linguistic levels. Figures 1 and 2 show the prominence of the same sentence, rated on the word level by one group of subjects (Fig. 1) and on the syllable level by another group of subjects (Fig. 2). For instance, the word prominence of "Berlin" is much higher than the syllable prominence of "lin", which carries lexical stress in the word "Berlin". There is a significant difference in the first word "In", probably due to the big difference in the prominence of the neighbour. Effects of these types were very frequent. In general, word prominence tends to be stronger than the prominence of the pertinent lexically stressed syllable. Often, units show differences in prominence ratings caused by the prominence variability of the context. Table 1 shows that the correlations between acoustic features and the prominence ratings are greater for the ratings on the word level. Combined with the finding of lower costs (cf. [18]) for the rating on the word level, one can conclude that rating on the word level is easier for naïve listeners than on the syllable level.

4. Normalisation

The normalisation of prominence ratings is reported rarely (but see [4, 24]). In [23] different approaches of normalisation of prominence ratings were evaluated on one data set. The study found that normalisation of prominence ratings significantly improves the correlation between prominence ratings and acoustic correlates of prominence. We applied a normalisation to the data of [17] and [18] by using the z-transformation, which is described by equation 1.

Table 1: *Acoustic correlates (Pearson cc) for the raw data of study [18]*

	Word prominence	Syllable prominence
Dauer	r= .69 ; $p < .001$	r= .41 ; $p < .001$
Maximum f0	r= .54 ; $p < .001$	r= .40 ; $p < .001$
Intensitt	r= .53 ; $p < .001$	r= .39 ; $p < .001$

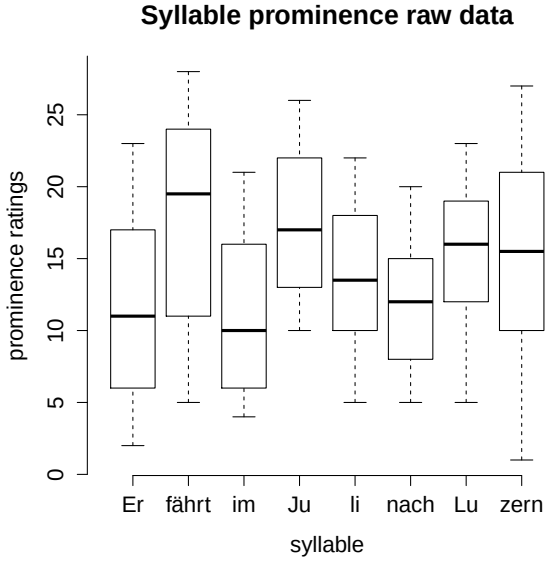

Figure 3: *Raw syllable prominence ratings of a sentence in study [18]. Er fährt im Juli nach Luzern - He will go to Luzern in July.*

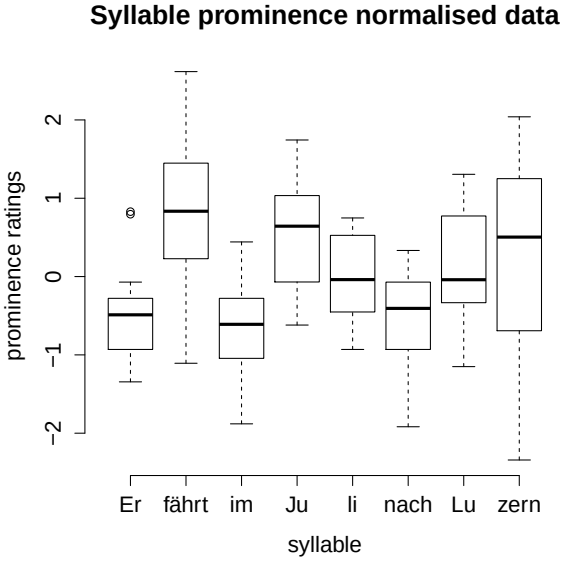

Figure 4: *Normalised syllable prominence ratings of a sentence in study [18]. Er fährt im Juli nach Luzern - He will go to Luzern in July..*

$$Z_n = \frac{R_n - \mu}{\sigma} \tag{1}$$

R_n is the prominence on the syllable n, μ is the mean and σ the standard deviation of all ratings. All statistical calculations were carried out by means of R [22].

4.1. Results

We found several advantages in using normalised prominence ratings. Correlations between acoustics and prominence ratings improve in most cases in both data sets. Table 2 shows the correlations between acoustics and the normalised prominence ratings from [18]. In [17] the priming effect could only be replicated using the 31-point scale. We primed each group with a prominence pattern to a syntactic structure and measured whether a syllable that differed in the pattern of the one group was rated significantly different by the other group where the prominence of that syllable matched the priming pattern. After the normalisation the effect was replicated with the 31-point and the 11-point scale (see Table 3). The effect is not significant for the ratings obtained with the 4-point and continuous scale. Figure 3 shows the raw prominence ratings of a sentence from the data of [18]. The last syllable carries the lexical stress of the last word. One would expect this syllable to receive a higher prominence rating than the first syllable of the same word. However, we see that the last syllable shows large disagreement in the ratings and that the mean of the last syllable is below the mean of the first syllable. Figure 4 shows the normalised prominence ratings for the same sentence. It is evident that after normalisation the ratings are much more in line with the linguistic expectations.

Table 2: *Acoustic correlates (Pearson cc) for the normalised data reported in [18]*

	Word prominence	Syllable prominence
Duration	r= .70 ; p $<$ .001	r= .39 ; p $<$ .001
Maximum f0	r= .54 ; p $<$ .001	r= .43 ; p $<$ .001
Intensity	r= .54 ; p $<$ .001	r= .44 ; p $<$ .001

4.2. Discussion

The general findings from [17] and [18] are not changed by using normalised data. The normalisation shows some positive effects on the prominence profiles. The correla-

Table 3: *Results of the priming in [17] using the raw data and normalised data with the four different scales. The Wilcoxon test shows whether the manipulated syllable is rated significantly different by the two groups of subjects.*

	4-point	11-point	31-point	continuous
Raw	W = 140.5 p = .49	W = 185.5 p = .46	W = 229 p $<$ .05	W = 143.5 p = .56
Z	W = 175 p = .34	W = 342 p $<$ .05	W = 229 p $<$ .05	W = 171 p = .39

tions between the acoustic features and the normalised prominence ratings are better after the normalisation for both data sets. This is in line with the findings in [23]. Moreover, prominence perception agrees better with linguistic expectations. As expected, normalisation compensates artefacts in the ratings, as shown in Figures 3 and 4. Since the effect of priming was preserved, these results give further support for the finding that priming of prominence pattern can be effective. After normalisation, the priming effect could be replicated with both the 11- and 31-point scale.

5. Conclusions

We conclude that the use of a scale with a larger number of levels provides interesting insights into the perception of prominence by naïve listeners. Contrary to ratings with a binary scale, one can observe both the perceived prominence profiles and inter-rater confidence. That is, subjects show a higher agreement on judging prominence of particular units, but units with high agreement are not necessarily the most prominent ones in an utterance. It will be interesting to investigate these variations in more detail. The ratings obtained with multilevel scales yield good correlations to the acoustics and facilitate the detection of rating differences. We did not find any disadvantages related to rating difficulties. The priming effect from [13] was replicated using the 11-point and 31-point scale. Its replication failed with the 4-point and the continuous scale.

Word prominence is easier to rate than syllable prominence for naïve listeners. Since the results of prominence retings obtained on the word and syllable level differ significantly, one should be careful when comparing results from studies using different levels of units for annotation.

We found that the normalisation of prominence ratings has several advantages. Prominence profiles are more in line with well-known acoustic correlates and show a better discrimination of rating differences. Furthermore, normalisation results in fewer artefacts in the prominence ratings. The effect that within a given type of unit the subjects agree more on particular units than on others does not disappear after normalisation. In sum, normalisation does not alter the main findings of [17, 18] but tends to make results more robust.

6. References

[1] Watson, D. G., Arnold, J. E. and Tanenhaus, M. K., "Tic Tac TOE: Eects of predictability and importance on acoustic prominence in language production", Cognition 106, 1548–1557, 2008.

[2] Wagner, P., "Wahrnehmung und Vorhersage deutscher Betonungsmuster", Universität Bonn. PhD-Thesis, 2002. Online: http://hss.ulb.uni-bonn.de/2002/0054/0054.htm, accessed on 29 Mar 2011.

[3] Fant, G und Kruckenberg, A., "Preliminaries tothe study of Swedish prose reading and reading style", STR-QPSR, 2/1989 KTH, Stockholm, 1–83, 1989.

[4] Eriksson, A., Grabe E. und Traunmüller, H., "Perception of syllable prominence by listeners with and without competence in the tested language". Proceedings Speech Prosody 2002, Aix-en-Provence, 275–278, 2002.

[5] Gussenhoven, C., Repp, B. H., Rietfeld, A. and Terken, J., "The perceptual prominence of fundamental frequency peaks", Journal of the Acoustical Society of America 102, 3009–3022, 1997.

[6] Gussenhoven, C. and Rietveld, T., "On the speaker-dependence of the perceived prominence of F0 peaks," Journal of Phonetics 26, 371–380, 1998.

[7] Kochanski, G., Grabe, E., Coleman, J., and Rosner, B., "Loudness predicts prominence: fundamental frequency lends little". Journal of the Acoustical Society of America 118, 1038–1054, 2005.

[8] Goldman, J.-P., Auchlin, A., Roekhaut, S., Simon, A. C., and Avanzi M., "Prominence perception and accent detection in French. A corpus-based account", Proceedings of Speech Prosody 2010, Chicago, 2010.

[9] Widera, C., Portele, T. and Wolters, M., "Prediction of word prominence", Proceedings of Eurospeech 1997, 999–102, 1997.

[10] Tamburini, F. "Automatic prosodic prominence detection in speech using acoustic features: an unsupervised system", Proceedings of Eurospeech 2003, 129–132, 2003.

[11] Wang, D. and Narayanan, S.; , "An Acoustic Measure for Word Prominence in Spontaneous Speech." IEEE Transactions on Audio, Speech, and Language Processing, vol.15, no.2, 690–701, 2007.

[12] Wagner, P., "Great Expectations - Introspective vs. Perceptual Prominence Ratings and their Acoustic Correlates", Proceedings of Interspeech 2005, Lisbon, 2381–2384, 2005.

[13] Arnold, D., Wagner, P. and Möbius, B., "The effect of priming on the correlations between prominence ratings and acoustic features", Proceedings of Speech Prosody 2010, Chicago, 2010.

[14] Streefkerk, B., "Prominence - Acoustic and lexical/syntactic correlates", Utrecht: LOT, 2002.

[15] Cole, J., Mo, Y. and Hasegawa-Johnson, M., "Signal-based and expectation-based factors in the perception of prosodic prominence", Laboratory Phonology 2010, 1:2, 425–452, 2010.

[16] Mo, Y., Cole, J. and Lee, E.-K. "Naïve listeners prominence and boundary perception", Proceedings of Speech Prosody 2008, Campinas, 2008.

[17] Arnold, D., Wagner, P., and Möbius, B. "Evaluating different rating scales for obtaining judgments of syllable prominence from naïve listeners", Proceedings of ICPhS 2011, Hong Kong, 2011.

[18] Arnold, D., Möbius, B., and Wagner, P., "Comparing word and syllable prominence rated by naïve listeners", Proceedings of INTERSPEECH 2011, Florence, 2012.

[19] Jensen, C. and Tøndering, J. , "Choosing a Scale for Measuring Perceived Prominence", Proceedings of Interspeech 2005, Lisbon, 2385–2388, 2005.

[20] Turk, A. E. and Sawusch, J. R., "The processing of duration and intensity cues to prominence", Journal of the Acoustical Society of America 99, 3782–3790, 1996.

[21] Grover, C., Heuft, B. und van Coile, B., "The reliability of labeling word prominence and prosodic boundary strength", Proceedings of the ESCA Workshop on Intonation, Athens, 165–168, 1997.

[22] R Development Core Team. "R: A language and environment for statistical computing.", R Foundation for Statistical Computing, Vienna, 2012.

[23] Sappok, C. and Arnold, D., "On the Normalization of Syllable Prominence Ratings", Proceedings Speech Prosody 2012, Shanghai, 2012.

[24] Liljencrants, J., "Judges of prominence", Fonetik 99: Proceedings from the Twelfth Swedish Phonetics Conference, Göteborg, 101–107, 1999.