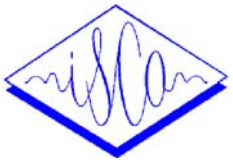

Speech Prosody 2004
Nara, Japan
March 23-26, 2004

Measuring the Reliability of Manual Annotations of Speech Corpora

Ulrike Gut[*] *& Petra Saskia Bayerl*[†]

[*]Albert-Ludwigs-University, Freiburg, Germany
[†]Justus-Liebig-University, Giessen, Germany
ulrike.gut@anglistik.uni-freiburg.de, petra.s.bayerl@psychol.uni-giessen.de

Abstract

The quality of manual annotations of speech corpora depends on the ability of human annotators to cope with phonetic and prosodic coding schemas such as SAMPA and ToBI. It has been proposed widely that an acceptable amount of reliability among and within individual annotators is impossible to achieve. In this paper, we present an extensive evaluation of annotator reliability in a multilevel phonetically annotated speech corpus, using two methods for measuring annotator reliability. The results show that manual annotations can be very reliable, but that reliability is correlated with the complexity of the coding schema.

1. Introduction

A growing demand for both technological applications and the theoretical development of models of spoken language has generated a multitude of annotated speech corpora. Despite efforts to automatize annotations [5, 22, 7, 2, 1, 6] manual annotations supported by various graphic and acoustic tools still play an important role in the compilation of these corpora. The quality of such manual annotations has been criticized on the following points:

- **implicit incoherence:** the manual labeling procedure is incoherent due to human variability in perceptual capabilities and other factors [4]. Intra-annotator reliability can never be perfect [6]
- **lack of consensus on coding schema:** manual annotations reflect the variability of the interpretation and application of the coding schema by the annotators [4]
- **annotator characteristics:** individual characteristics of coders such as familiarity with the material, amount of former training, motivation and interest and fatigue-induced errors influence the quality of annotations [19]

So far, no widely accepted methodology for the measurement of the quality of manually annotated speech data exists [24]. Various methods have been proposed including pairwise comparisons [19, 10, 21], Cohen's kappa and Pearson's chi-square [21].

The goal of this paper is to present two approaches for measuring reliability of manual phonetic and prosodic annotations and to find factors of the degree of reliability. In Section 2, the LeaP corpus of non-native speech, the annotation process of the LeaP data and the training of the annotators are described. In Section 3, the reliability of the annotations are presented in terms of both inter-annotator agreement and intra-annotator agreement. In Section 4 the results of this evaluation are discussed.

2. Annotation in the LeaP corpus

2.1. The LeaP corpus

The LeaP corpus was collected in the LeaP (Learning Prosody) project[1], which is concerned with the acquisition of prosody by non-native speakers of German and English. The aims of the project include both the phonetic and phonological description of non-native prosody and the exploration of learner variables that influence the acquisition process. Data was collected from different groups of speakers: learners before and after a period abroad, before and after a four-month prosody training course, especially advanced learners who are hardly distinguishable from native speakers, and learners with different levels of competence. A quasi-experimental study was carried out that compared a treatment group of students taking part in a theoretical and practical training course in prosody with a control group. The data collected in the training courses include measurements of perception as well as production. The recordings consist of readings of nonsense word lists and three different speech styles:

- Readings of a short story (about 2 minutes).
- Re-tellings of the same story (between 2 and 5 minutes).
- Interviews (between 10 and 30 minutes).

2.2. Annotation

The manual annotation of the LeaP data was carried out using ESPS/waves+ and Praat and comprises six different tiers:

1. On the phrase tier, speech and non-speech intervals are transcribed. The interviewee's speech is divided into intonational phrases.
2. On the words tier, the beginning and end of words in the speech of the interviewee are transcribed.
3. On the syllable tier, the beginning and end of syllables are marked.
4. On the segments tier, all vocalic and consonantal intervals plus the intervening pauses are annotated.
5. On the tones tier, pitch accents and boundary tones are annotated.
6. On the pitch tier, the initial high pitch, the final low pitch and intervening high peaks and low valleys are annotated.

In addition, two tiers were added automatically with part-of-speech coding and an assignment of the words to lemmata. For a recording of about one minute length, on average, 3000 events are annotated. The entire corpus consists of 359 files

[1]http://leap.lili.uni-bielefeld.de

annotated in this fashion and includes a total of 131 different speakers with 32 different native languages as well as 18 recordings of native speakers. The total amount of recording time is more than 12 hours.

2.3. Annotation schemas

Each of the six tiers uses different annotation schemas, most of which were developed for the specific purpose of the corpus. They all vary in complexity, some involving only a small number of different categories (e.g. the schema for the pitch tier), some a large one (e.g. the annotation schema on the tones tier). Only for a few of these annotation schemas annotator agreement studies have been carried out.

On the phrase tier, a total of nine categories can be annotated: Speech (by the interviewee), interrupted phrases, unfilled pauses, noise, breath, laughter, hesitation phenomena, elongated phonemes and speech by the interviewer. Speech by the interviewee is divided into intonational phrases, whose delimitation can be marked by final syllable lengthening, an intonational boundary tone and a following pause. The concept of phrases is similar to that of the ToBI break indices 3 and 4. For these, interrater reliabilities of between 67% and 86%, measured in pairwise agreement, have been reported [19, 10, 21].

On the words tier, speech is transcribed orthographically, allowing no capital letters in either German or English. Transcriptions of cliticizations such as "aren't" are possible.

On the syllable tier, syllables are transcribed in SAMPA [23]. For the transcription of consonants with SAMPA, a pairwise agreement of 94.8% on average was found by [24]. [25] reported Cohen kappa values of between 0.49 and 0.73 for pairs out of nine experienced phoneticians judging whether a phone was present or not. The determination of syllable boundaries is based on auditory criteria which allow for resyllabification processes in spoken language [8]. To our knowledge, there are no previous studies on annotator agreement in terms of syllable boundaries.

On the segments tier, vocalic and consonantal speech intervals are transcribed where vowels and postvocalic semi-vowels are considered vowels and plosives, fricatives, nasals, approximants, affricates, prevocalic semivowels, laterals, trills and retroflexes are considered consonants. The determination of the segment boundaries is supported by a broad band spectrogram and carried out following phonetic standard criteria [17]. [24] report an average deviation of segment boundaries for their annotators of between 7 and 15 ms, the last found for segment boundaries between two adjacent vowels. Overall, 93% of all transcribed segment boundaries lie within less than 15 ms.

For the tone tier, a modified version of EToBI and GToBI [19, 9] was developed in the project. In total, 14 different types of pitch accents (including downstep and upstep and all possible compound pitch accents) and 14 different types of boundary tones can be used. The language-specific variants of ToBI have become the standard transcription system of the auto-segmental metrical model of intonation and a number of studies have been carried out testing its reliability between annotators, varying in the number of annotators, their heterogeneity in terms of training and speech material used [19, 9, 21, 18]. Pairwise agreement of between 61% and 71% for pitch accents and 84% and 95% for boundary tones were found.

On the pitch tier, four categories of pitch height can be annotated, following [16].

2.4. Annotator characteristics

A total of six annotators worked in the project. Only annotator 4 had had any prior experience (about 5 years) in phonetic and prosodic annotation. Annotators 1, 2, 3 and 4 had an intensive (about 8 hour/week) training course in annotation for three months at the beginning of the project. In this training phase, criteria for the categories on the phrase, syllable, word and segments tiers were established and the annotations were discussed both individually and in the group. Annotator 5 joined the project later and had only a short training phase of about three weeks. Annotators 1, 3, 4 and 6 were trained in ToBI transcription, both with the help of the GToBI training web pages and in an intensive two-day workshop with an expert in GtoBI.

Table 1: *Annotator characteristics in terms of tiers annotated and training*

Annotator	**Tiers Annotated**	**Training**
1	all	3 months training, ToBI training
2	phrase, words, syllables, segments	3 months training
3	all	3 months training, ToBI training
4	all	5 years experience, 3 months training, ToBI training
5	phrase, words, syllables, segments	3 weeks training
6	tones	ToBI training

3. Methodology

In general, reliability can be defined as "the complex property of a series of observations or of the measuring process that makes it possible to obtain similar results if the measurement is repeated" [11]. Since the degree of agreement can be considered an indicator for reliability, two measurements were taken in this study: the agreement between all annotators annotating the same speech recording (inter-annotator agreement) and the agreement for each annotator with himself annotating the same speech recording twice (intra-annotator agreement). The first is calculated as pairwise agreement between two different raters, which allows conclusions about the stability of annotations; the second indicates the reproducability of annotations [12]. For deeper insights into the nature of agreement, sources for disagreement have been investigated.

3.1. Data

For the calculation of inter-annotator agreement all annotators annotated one recording of a 268-word re-telling of the story. Annotators 1, 3 and 4 annotated all tiers, annotators 2 and 5 four tiers and annotator 6 one tier as described in Table 1. For the measurement of intra-annotator agreement annotators 1 to 5 annotated one speech recording twice with an interval of nearly two years between the first and the second annotation. In case of the sixth tier (pitch) only three, and at the fifth tier (tones) only four persons annotated the data twice. The data comprised between 129 (phrase) and 988 (vowels) segments. Labels for the same events were determined according to their time-stamps.

Table 2: *Inter-annotator agreement for different tiers [kappa values]*

	Coder Pairs														
Tier	1-2	1-3	1-4	1-5	1-6	2-3	2-4	2-5	3-4	3-5	3-6	4-5	4-6	**M**	**R**
1. Phrase	.77	.78	.73	.69	–	.75	.82	.66	.76	.73	–	.61	–	.73	.21
2. Words	.95	.96	.96	.96	–	.93	.92	.92	.99	.93	–	.94	–	.95	.07
3. Syllables	.28	.31	.37	.23	–	.25	.30	.15	.30	.22	–	.23	–	.26	.22
4. Segments	.99	.99	.99	.98	–	.98	1.00	.99	.99	.97	–	.99	–	.99	.03
5. Tones	–	.26	.37	–	.26	–	–	–	.32	–	.36	–	.38	.33	.12
6. Pitch	–	.79	.94	–	–	–	–	–	.88	–	–	–	–	.87	.18

M – Mean; R – Range

3.2. Measuring agreement

Intra- and inter-annotator agreement were both calculated by unweighted kappa (κ) as defined by [3]. Overall agreement was measured as the mean of the kappa values on one tier.

3.3. Investigating sources of disagreement

Low agreement can be due either to differences in the interpretation of labels or categories, or to differences in the probability of category use. These sources of disagreement can be analysed by means of the consecutive use of (a) the Maxwell-Stuart test for inequality of marginals [14, 20] and (b) the McNemar test [15]. If the Maxwell-Stuart test is significant, different interpretations of labels must be assumed. The McNemar test then indicates which labels are the problematic ones.

4. Results

Table 2 shows that the inter-annotator agreement differs considerably for different tiers. Mean kappas for syllables and tones are as low as .26 and .33, indicating only fair agreement, whereas words, segments, and pitch have values of .95, .99, and .87, respectively, indicating almost perfect agreement [13]. The mean agreement of .73 on the phrase tier is slightly higher than that reported by [21] for ToBI break indices, although those two annotation schemas are of course not directly comparable. Kappa values for ToBI tones are much lower in this study than in [21]. Intra-annotator agreement presents similar results as inter-annotator agreement (Table 3). Here, syllables also achieved the lowest ($M = .39$), segments ($M = .97$) and words ($M = .96$) the highest kappa values.

Table 3: *Intra-annotator agreement for different tiers [kappa values]*

	Coder						
Tier	1	2	3	4	5	**M**	**R**
1. Phrase	.61	.35	.62	.87	.67	.62	.52
2. Words	.91	.92	.99	.99	.97	.96	.08
3. Syllables	.34	.32	.38	.57	.36	.39	.25
4. Segments	.98	.99	.95	.98	.97	.97	.04

M – Mean; R – Range

The consistent differences between the tiers in both measurements of agreement show that the reliability of manual annotation is mostly influenced by the complexity of the annotation task: the number of categories and kappa values are significantly correlated ($r = -.65;\ p < .001$). Not surprisingly, values are lowest for the syllable tier, where annotators were required to carry out two types of annotation: a segmentation into syllables as well as a transcription in SAMPA. Results for intra-annotator agreement indicate that experience with annotations may positively influence the reproducibility in difficult tasks, e.g. annotation of the syllables by annotator 4. Lack of experience might also be the reason for the lower kappa values on the tones tier achieved by the annotators in this study compared to the extensively trained annotators in [21]. However, a more detailed study is needed for a substantiation of these first results.

An interesting point to note is that the ranges of kappa values (maximum minus minimum) for inter-annotator agreement are much larger when the annotation task is difficult. In easier cases with high inter- and intra-annotator agreement, ranges are smaller. All in all, ranges in both cases are relatively small indicating that the degree of agreement is attributable primarily to the annotated tier. This can also be proven statistically, since kappa values for annotation tiers were significantly different for inter-annotator agreement ($\chi^2 = 44.82;\ df = 5;\ p < .001$) as well as for intra-annotator agreement ($\chi^2 = 15.43;\ df = 3;\ p < .05$), whereas annotators (intra-annotator agreement, $\chi^2 = 1.28;\ df = 4;\ p = .87$) as well as annotator pairs (inter-annotator agreement, $\chi^2 = .58;\ df = 11;\ p = 1.00$) showed no significant differences in annotation quality.

As investigations of agreement patterns between annotators showed, actual disagreement was caused only by a small number of categories. At the phrase tier, for instance, the two categories *noise* and *pause* were frequently confused with each other. On the tone tier 20 of the 46 categories were significantly different, but only five of them were relevant for three or more of the six annotator pairs. As expected, for the sixth tier (pitch) with its very high levels of agreement, no significant differences were found. Data for the syllables and words tier with approximately 450 and 268 different categories, respectively, was too sparse for statistical testing.

5. Conclusions

The object of this study was to assess the quality of manual annotations and to find factors that influence inter- and intra-annotator agreement. It was shown that almost perfect agreement between annotators is possible, but that agreement is correlated to the complexity of the annotation task. The higher the number of different categories in an annotation scheme, the lower the agreement. However, the results show that the number of annotation schema categories which lead to confusions among annotators is often relatively small so that an improvement of annotation reliability can be achieved fairly easily by carrying out systematic error analyses as suggested here and by changing the annotation schema accordingly. The results furthermore suggest that experience with annotation has a major influence on reliability.

6. References

[1] N. Braunschweiler. 2003. ProsAlign – the automatic prosodic aligner. 3093 – 3096.

[2] I. Bulyko and M. Ostendorf. 2002. A bootstrapping approach to automating prosodic annotation for limited-domain synthesis. In *Proc. IEEE Workshop on Speech Synthesis*, September 2002.

[3] J. Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37 – 46.

[4] P. Cosi, D. Falavigna, and M. Omologo. 1991. A preliminary statistical evaluation of manual and automatic segmentation discrepancies. 693 – 696.

[5] P. Cosi. 1993. SLAM: Segmentation and labeling automatic module. 88 – 91.

[6] C. Cucchiarini and H. Strik. 2003. Automatic transcription agreement: An overview. 347 – 350.

[7] J. Garica, U. Gut, and A. Galves. 2002. Vocale - A semi-automatic annotation tool for prosodic research. 327 – 330. Laboratiore Parole et Langage.

[8] H. Giegerich. 1992. *English phonology*. Cambridge University Press, Cambridge.

[9] M. Grice, S. Baumann, and R. Benzmüller. 2002. German intonation in autosegmental-metrical theory. In S.-A. Jun, editor, *Prosodic Typology*. Oxford University Press.

[10] M. Grice, M. Reyelt, M. Benzmüller, J. Mayer, and J. Batliner. 1996. Consistency in transcription and labeling of German intonation with GToBi. 1716 – 1719.

[11] E. Hollnagel. 1993. *Human Reliability Analysis Context and Control*. Academic Press.

[12] K. Krippendorff. 1980. *Content analysis: An introduction to its methodology*. Sage Publications.

[13] J. Landis and G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33:159–174.

[14] A. Maxwell. 1970. Comparing the classification of subjects by two independent judges. *British Journal of Psychiatry*, 116:651 – 655.

[15] Q. McNemar. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12:153 – 157.

[16] D. Patterson. 2000. *A linguistic approach to pitch range modelling*. PhD thesis, Edinburgh University.

[17] G. Peterson and I. Lehiste. 1960. Duration of syllable nuclei in English. *Journal of the Acoustical Society of America*, 32(6):693 – 703.

[18] J. Pitrelli, M. Beckman, and J. Hirschberg. 1994. Evaluation of prosodic transcription labeling reliability in the ToBI framework. 123 – 126.

[19] K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg. 1992. ToBI: A standard for labeling English prosody. In *Proceedings of the 1992 International Conference on Spoken Language Processing*, 867 – 870.

[20] A. Stuart. 1955. A test for homogeneity of the marginal distributions in a two-way classification. *Biometrika*, 42:412 – 416.

[21] A. Syrdal and J. McGory. 2000. Inter-transcriber reliability of ToBI prosodic labeling. *Proceedings of ICSLP, Bejing*.

[22] A. Vorsterman, J.-P. Martens, and B. van Coile. 1996. Automatic segmentation and labelling of multi-lingual speech data. *Computational Linguistics*, 19(4):271 – 293.

[23] J. Wells, W. Barry, M. Grice, A. Fourcin, and D. Gibbon. 1992. Standard computer-compatible transcription. Technical report, Tech. Rep. No. SAM Stage Report Sen.3 SAM UCL-037.

[24] M. Wesenick and A. Kipp. 1996. Estimating the quality of phonetic transcriptions and segmentions of speech signals. *Proceedings of ICSLP*.

[25] C. M. Wester, J. Kessens and H. Strik. 2001. Obtaining phonetic transcriptions: A comparison between expert listeners and a continuous speech recognizer. *Language and Speech*, 44(3):377 – 403.