Reliability Report:
Internship at UvA Speech Lab

Name: Yuying Zhu
Student Number: 14452685
Supervised by: Dr. A.T. (Titia) Benders, M.H. (Marloes) Roosingh

Table of Contents

1. Overview

1.1 Overall goal

The reliability analysis aims to assess the consistency of annotations made by two coders on the same set of audio files for a pilot study in the research project titled "Perception-Production Link in Child Language", led by Dr. A. T. Benders. The study aims to investigate the perception and production of speech prosody of English-speaking children in Australia. By comparing their coding results, this analysis evaluates the reliability of the existing annotation schemes and identifies potential areas for enhancement. The ultimate goal is to improve the quality and effectiveness of future annotations.

1.2 Theoretical background

As reported by Yuen et al. (2022), a study by Wheeldon & Lahiri (1997) discovered that in a delayed production task for Dutch sentences, the reaction times corresponded to the number of Prosodic Words (PW) rather than the number of Orthographic Words or syllables. This finding supports the idea that PW, along with the prosodic cliticization of function words, play a crucial role in as the planning unit in speech production. This was also attested to be true for English by Wynne et al. (2018, cited in Yuen et al., 2022).

Previous studies have established that English-speaking two-year-olds frequently omit unstressed syllables followed by the syllable with primary stress when producing polysyllabic words (Allen & Hawkins, 1980; Klein, 1981; Gerken, 1994; cited in Carter & Gerken, 2002). However, a study by Carter & Gerken (2002) found that children do not entirely delete these weak syllables. Instead, some prosodic trace of the omitted syllable remains, in the form of a significantly lengthened duration.

According to Cai et al. (2024), English articles, which act as prosodic clitics, can attach to a nearby lexically stressed content word. However, the direction in which monosyllabic articles are cliticized by children remains undetermined. Previous studies have found evidence supporting both leftward and rightward cliticization in children's speech production (see *Figure 1*). Therefore, comparing the durations of different speech segments in children's cliticized and uncliticized utterances may provide insight

on the inner mechanism of their perception and production of speech prosody.

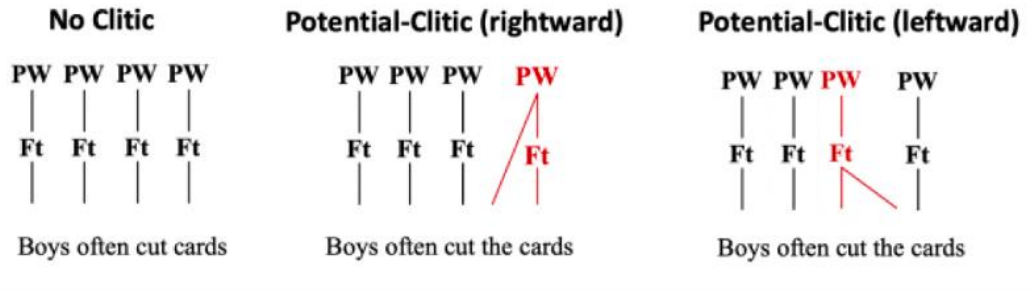

Figure 1: two potential directions of Potential-Clitic (Cai et al., 2024)

1.3 Data collection and annotation

During the data collection procedure, participants were asked to engage in an elicited imitation task with the aid of visual stimuli conducted in the form of conversations like:

Experimenter: “Who sees a wombat?”

Participant: “Minnie sees a wombat.”

The second half of the sentence produced by the participant—specifically, the underlined verb phrase (VP) that consists of a verb, a noun, and occasionally an article—constitutes the primary focus of this study. In particular, the durations of various segments of the constituents within the VP are of interest, such as the duration from the onset to the end of the article, or from the onset to the end of a third-person singular suffix. These durations are marked by specific landmarks that indicate the start and end points of the relevant speech segments. For example, the landmark “ssv” marks the beginning of the stressed vowel in the verb, while “su.v” and “eu.v” denote the start and end points of the unstressed vowel found either in the noun or the article. An example of an annotated utterance is illustrated in *Figure 2*. The landmarks (“points” in *Praat*) on tiers 2 to 4 were those involved in the annotation task.

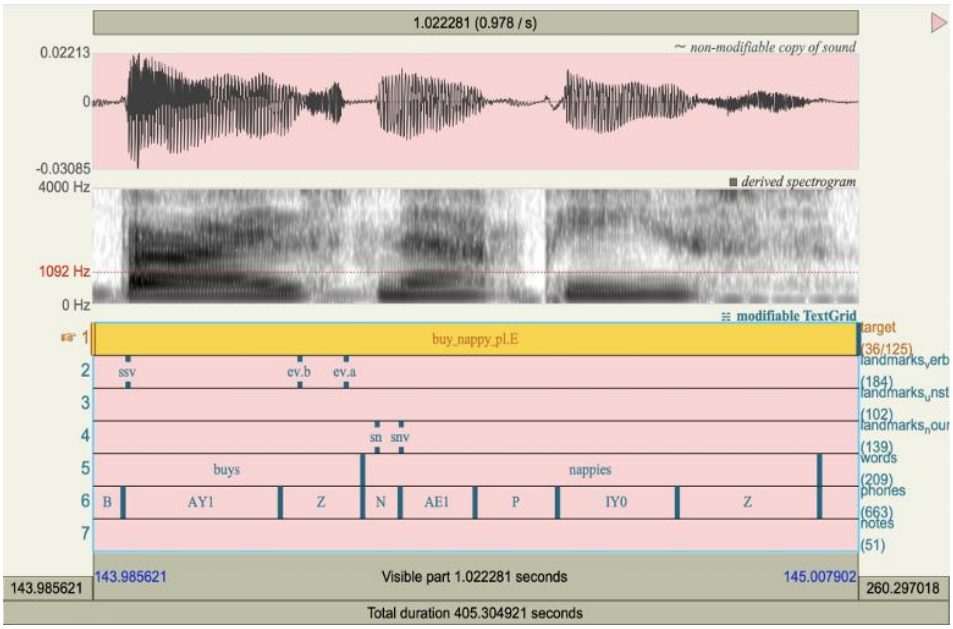

Figure 2: An annotated utterance “buys nappies”

Initially, these landmarks were generated and aligned automatically using *Montreal Forced Aligner* (McAuliffe et al., 2017). However, the timing accuracy of these landmarks was insufficient for the study’s requirements, which necessitated precision at the millisecond level. Therefore, to accurately locate the speech segments of interest, it was essential to manually adjust the positions of the automatically generated landmarks, using spectrograms as the primary reference and occasionally consulting the waveform for additional clarity.

2. Data Collection and Data Processing

2.1 Re-coding and utterance selection

The reliability analysis was carried out based on the annotations made by a former intern at the UvA Speech Lab on 8 recordings of 4 female adult speakers (i.e. files numbered 82051 to 82112). Data collection for the reliability analysis was conducted through re-coding approximately 20% of the utterances annotated by the original coder in *Praat* (Boersma & Weenink, 2024). In this report, the original coder will be referred to as *Coder 1*, and the re-coder as *Coder 2*.

To ensure variation in the utterances involved in the reliability analysis and an even distribution of the utterances among speakers, a combination of controlled and randomized selection methods was used to choose the utterances for re-coding. For a more detailed description of the re-coding scheme, see *Appendix 1*.

A total of 79 utterances were re-coded for the reliability analysis. For the re-coding log data (i.e., the re-coded utterances), see *Appendix 2*.

2.2 Data processing

The raw data extracted from *Praat* was processed with *Python* (Van Rossum & Drake, 1995) scripts. More detailed explanations on the scripts used for the data processing can be found in *Appendix 3*. For the raw data and processed data, see *Appendix 4*.

3. Results

3.1 Overview

A total of 554 data points (i.e., annotated landmarks) were involved in the reliability analysis. For an overview of the inter-coder consistency of each annotated boundary, see *Appendix 4*.

3.2 Differences between coders

The descriptive statistics for the absolute differences and the quantiles of differences and absolute differences between coders are as shown in *Table 1* and *2*. *Figure 3* displays the distribution and quantiles of both the differences and absolute differences between the coders. For better visualization, 29 data points (5% of the total) with extreme values exceeding ±20ms were trimmed.

The visualizations and descriptive statistics were generated using *R* (R Core Team, 2013). For the *R* scripts used, see *Appendix 3*.

(ms)	Mean	SD	Median	Range
Absolute Differences	5.44	8.80	2.42	90.43

Table 1: Descriptive statistics for the absolute differences between coders

(ms)	25%	50%	75%
Differences	-3.93000	0.10950	1.55125
Absolute Differences	0.83875	2.41600	6.71100

Table 2: Quantiles of differences and absolute differences between coders

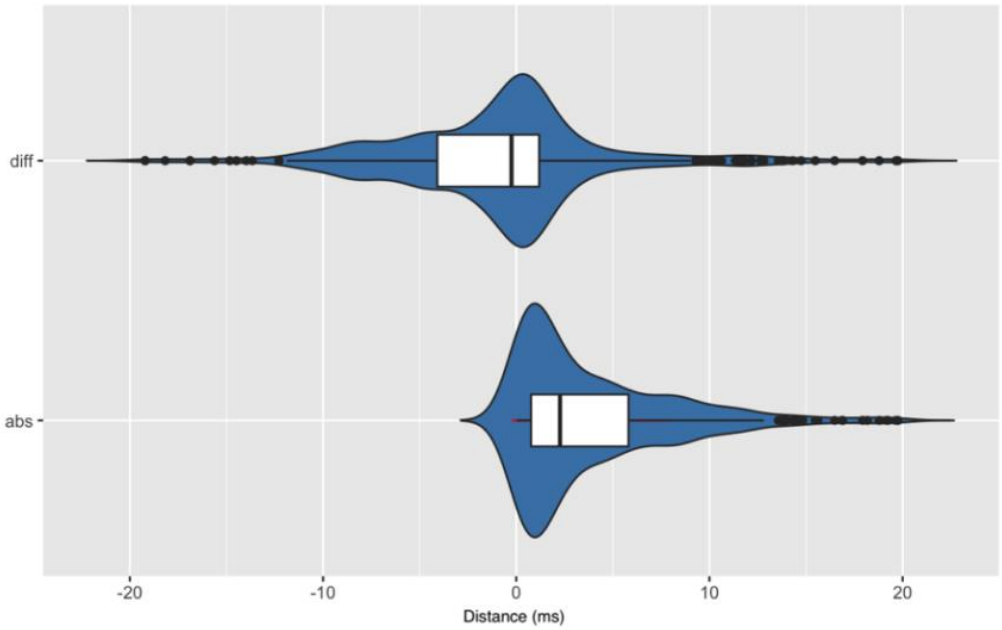

Figure 3: Distribution and quantiles of differences and absolute differences

4. Inter-coder Consistency

4.1 Criterion for consistency

The annotations to a specific landmark by the two coders are deemed consistent if the two annotations are less than 5ms (=0.005s) apart from each other. Likewise, the lengths of corresponding intervals are deemed consistent if their mean difference is below 5ms.

4.2 Consistency by percentages

An overview of the inter-coder consistency by percentages is provided in *Table 3*.

	Landmark	> 5ms (inconsistent)	< 5ms (consistent)	Number of tokens
Total	-	178 (32%)	376 (68%)	554
Verb	***ssv***	7 (9%)	72 (91%)	79
	ev.b	16 (20%)	63 (80%)	79
	ev.a	36 (46%)	42 (54%)	78
Total		59 (31%)	177 (69%)	236
Unstressed	***su.c***	4 (27%)	11 (73%)	15
	su.b1	2 (18%)	9 (82%)	11
	su.v	14 (25%)	31 (75%)	45
	eu.v	21 (48%)	23 (52%)	44
Total		41 (33%)	74 (67%)	115
Noun	***sn.c***	23 (42%)	32 (58%)	55
	sn.b1	7 (14%)	42 (86%)	49
	sn.b2	3 (60%)	2 (40%)	5
	snv	28 (39%)	44 (61%)	72
	sn	17 (77%)	5 (23%)	22
Total		78 (38%)	125 (62%)	203

Table 3: Inter-coder consistency by percentage

However, as pointed out by Cohen (1960), merely stating the overall agreement percentage is insufficient for assessing reliability (Lane et al., 2023). Therefore, more reliable statistical methods for consistency analysis should be considered.

4.3 Assessment of inter-coder consistency: Cohen's kappa

To achieve a more accurate assessment of the consistency between coders, Cohen's kappa coefficient (Cohen, 1960) was employed for the purpose of the reliability analysis. The Cohen's kappa, as shown in the equation below, is a specialized measurement of reliability that accounts for chance agreement. In this context, P_O represents the observed agreement, while P_C denotes the expected agreement due to random chance.

$$\text{kappa} = \frac{P_O - P_C}{1 - P_C}$$

To note, the criteria for interpreting the Cohen's kappa coefficient vary quite significantly across different studies. The two interpretations of Cohen's kappa

utilized in this analysis are detailed in *Appendix 5*. In general, a kappa value above 0.60 indicates a moderate to strong level of agreement.

4.3.1 Data processing for kappa

Since Cohen' s kappa is only applicable to categorical data, the continuous outcomes obtained in Section 2.2 were converted into categorical data by assigning each data point a binary value, for instance, either 0 or 1. Specifically, for each annotation made by Coder 1, a value of 0 or 1 was assigned randomly. For the corresponding annotation made by Coder 2, the same value was assigned if the difference between the two annotations was less than 5ms; otherwise, the opposite value was assigned. The scheme for converting continuous data into categorical data is outlined in *Table 4*.

	Coder 1	Coder 2
consistent	0	0
	1	1
inconsistent	0	1
	1	0

Table 4: Scheme of converting continuous data into categorical data

To assess the effectiveness of this data processing scheme, Cohen's kappa was calculated on the same data set 50 times, each time using a different set of randomized binary values. It was observed that while the kappa results exhibited significant variability for smaller data sets, the kappa values derived from larger sample sizes with more than 50 tokens were noticeably more stable. This suggests that the data processing method exhibits a level of effectiveness when applied to sufficiently large data sets.

The data processing and calculation of Cohen' s kappa was performed using *Python* scripts. Details of the scripts utilized can be found in *Appendix 3*.

4.3.2 Kappa results: landmarks (points)

The Cohen's kappa coefficient and level of agreement for the exact location of landmarks (i.e. points) are as shown below:

(1) Cohen's kappa: overall

An overview to the Cohen's kappa coefficient and level of agreement for all landmarks involved in the reliability analysis is provided in *Table 5*. Overall, the annotations made by the two coders exhibit minimal to fair levels of consistency.

	Landmark	Cohen's kappa	Level of Agreement (McHugh, 2012)	Level of Agreement (Rafieyan, 2016)
Total	-	0.35	minimal	fair

Table 5: Cohen's kappa overall

(2) Kappa by landmarks and word types

The Cohen's kappa by landmarks and word types are as shown in *Table 6*. Among all landmarks, only *ssv* demonstrates a strong level of consistency between coders. *su.b1*, *sn.b1*, and *ev.b* show moderate to substantial levels of agreement, while the remaining landmarks display minimal or no inter-coder consistency.

Word Type	Landmark	Cohen's kappa	Level of Agreement (McHugh, 2012)	Level of Agreement (Rafieyan, 2016)
Verb		0.49	weak	**substantial**
	ssv	0.82	**strong**	**almost perfect**
	ev.b	0.59	weak	**moderate**
	ev.a	0.055	none	slight
Unstressed		0.27	minimal	fair
	su.c	0.45	weak	moderate
	su.b1	0.62	**moderate**	**substantial**
	su.v	0.33	minimal	fair
	eu.v	0.042	none	slight
Noun		0.24	minimal	fair
	sn.c	0.16	none	fair
	sn.b1	0.71	**moderate**	**substantial**
	sn.b2	-0.20	-	poor
	snv	0.22	minimal	fair
	sn	-0.42	-	poor

Table 6: Cohen's kappa by landmarks

(3) Cohen's kappa by presence/absence of article

A preliminary comparison was conducted on the exact locations of landmarks within utterances with and without articles. On top of that, comparable utterances (for instance, "sees a bib" and "sees bibs") were selected out for a more precise analysis of the effect of the presence or absence of articles (i.e., clitics) on annotation accuracy. As shown in *Table 7*, the inter-coder consistency for utterances in all four conditions ranges from minimal to fair, and the kappa results indicate no significant difference between groups.

	Cohen's Kappa	Level of Agreement (McHugh, 2012)	Level of Agreement (Rafieyan, 2016)
+article (overall)	0.37	minimal	fair
-article (overall)	0.34	minimal	fair
+article (selected)	0.25	minimal	fair
-article (selected)	0.39	minimal	fair

Table 7: Overall Cohen's kappa by presence/absence of article

(4) Cohen's kappa by prosodic structure of the noun

A preliminary comparison was carried out on the precise locations of landmarks within utterances containing nouns with varying prosodic structures. As indicated in *Table 8*, only the utterances featuring nouns with a strong-weak (i.e., trochaic) structure demonstrated a moderate to substantial level of inter-coder consistency. In contrast, the annotations for the other three groups exhibited only weak to fair levels of agreement.

	Cohen's Kappa	Level of Agreement (McHugh, 2012)	Level of Agreement (Rafieyan, 2016)
s	0.34	weak	fair
sw	0.41	**moderate**	**substantial**
ws	0.30	weak	fair
wsw	0.32	weak	fair

Table 8: Cohen's kappa by prosodic structure of the noun

4.3.3 Kappa results: durations (intervals)

The Cohen's kappa coefficient and level of agreement for the durations of specific speech segments (i.e. intervals) are as shown below:

(1) Cohen's kappa for duration of unstressed vowel

The consistency of the durations of unstressed vowels, measured by the interval lengths between landmarks *su.v* and *eu.v*, is shown in *Table 9*.

The overall kappa value for the duration of unstressed vowels is below 0, indicating poor inter-coder consistency. Furthermore, the presence of an article in the utterance, namely, whether the unstressed vowel appears in the article or in the noun does not significantly affect the consistency of vowel duration.

	Cohen's Kappa	Level of Agreement (McHugh, 2012)	Level of Agreement (Rafieyan, 2016)
usv (overall)	-0.56	-	poor
usv (singular)	-0.54	-	poor
usv (plural)	-0.29	-	poor

Table 9: Cohen's kappa for duration of unstressed vowel

(2) Cohen's kappa for duration of 3SG suffix

Similarly, as shown in *Table 10*, the durations of the third-person singular suffix (measured by the interval lengths between landmarks *ev.b* and *ev.a*) displayed poor to negligible inter-coder consistency. Additionally, the presence of a following article had no significant impact on the consistency of its duration.

	Cohen's Kappa	Level of Agreement (McHugh, 2012)	Level of Agreement (Rafieyan, 2016)
3sg (overall)	-0.51	-	poor
3sg (singular)	-0.45	-	poor
3sg (plural)	-0.52	-	poor

Table 10: Cohen's kappa for duration of third-person singular suffix

5. Conclusions

5.1 Consistency of landmarks (points)

In summary, the landmarks that are attested to be reliable are: *ssv* (for verbs), *sn.b1* (for nouns) and *su.b1* (for unstressed). Landmark *ev.b* can also be considered as a potentially reliable landmark for future analysis if a landmark pinpointing the 3SG suffix is needed, though it is suggested that the reliability for *ev.b* should be further tested.

The results from the *sn* landmarks reveal notable issues upon closer examination. 17 of the 22 re-coded *sn* landmarks differ from the original coding by more than 5ms, resulting in a kappa value below 0. This may be because most of the phonemes marked with *sn* landmarks are liquids, nasals and approximants, whose quality varies greatly among speakers and often blends with adjacent sounds. This makes it especially difficult to establish universally applicable strategies for accurately identifying their onset.

Within the current data set, neither the presence of an article (clitic) nor the prosodic structure of the noun exhibit significant effect on annotation accuracy. No further conclusions can be drawn on the reliability of coding schemes regarding these features.

5.2 Consistency of durations (intervals)

More crucially, the durations of speech segments like unstressed vowels or third-person singular suffix are all of substantially low consistency, with kappa values below 0. As shown in *Table 11*, in combination with the consistency of landmarks obtained above, it is evident that the inconsistency is mainly caused by significant disagreements between coders regarding the ending points of these intervals, namely, landmarks *ev.a* and *eu.v*. Therefore, it is suggested that the annotation schemes to these two landmarks should be improved.

	Landmark	Cohen's kappa	Level of Agreement (McHugh, 2012)	Level of Agreement (Rafieyan, 2016)
start 3sg	***ev.b***	0.59	weak	moderate
end 3sg	***ev.a***	0.055	**none**	**slight**
start usv	***su.v***	0.33	minimal	fair
snd usc	***eu.v***	0.042	**none**	**slight**

Table 11: Cohen's kappa by start and end points of intervals

6. Discussions

6.1 Limitations

Several limitations were identified in the reliability analysis:

A general issue with the current analysis is the limited sample size. Since the current kappa analysis method is only applicable to sufficiently large data sets, the results for certain landmarks may not accurately reflect the actual consistency between coders in their annotations. For example, only five *sn.b2* landmarks were included in the reliability analysis, making the sample size too small to draw any valid conclusions. The sample size of other landmarks, for instance *sn.b2* and *su.b1* are also relatively small. It is suggested that the reliability of the annotations for these landmarks should be tested with a larger data set. Additionally, the overall sample size might also be limited, given that there are only two coders and eight files involved in the reliability analysis.

Another significant limitation is that the reliability design was finalized before I gained exposure to prosody in *Phonology*, which rendered the final data set unsuitable for a more in-depth analysis of prosodic features. I therefore recommend proposing a more systematic and theoretically grounded design for utterance selection for the re-coding.

6.2 Other statistical methods explored

Due to my limited math competence, the statistical methods employed in this analysis are likely to be of low effectiveness. Moreover, since the categories applied in the Cohen's kappa calculation were determined by a simple conditional judgment regarding whether the annotations made by the two coders are more than 5ms apart, it fails to capture more subtle differences in the original continuous data, namely, the extent of the disparity between the coders. Other statistical methods for calculating the reliabilities of continuous data like *Intraclass Correlation Coefficient* (Shrout and Fleiss, 1979, cited in Lane et al., 2023, see *Table 12*) and *Cronbach's alpha* (Cronbach, 1951, cited in Lane et al., 2023), as well as consistency analysis based on *Bland-Altman* plot (see *Figure 4*) were tested on the existing data set using *R* (R Core Team, 2013. For the *R* scripts, see *Appendix 3*). However, no valid results were yielded. It is suggested that other more reliable statistical methods should be considered for the reliability analysis.

Versions of Intraclass Correlation Statistics for Various Reliability Designs

Type of reliability study design	Raters fixed or random?	Version of intraclass correlation
Part A. Reliability of single rater		
Nested: *n* subjects rated by *k* different raters	Random	$ICC(1,1)=\frac{\hat{\sigma}_T^2}{\hat{\sigma}_T^2+\hat{\sigma}_w^2}$
Subject by rater crossed design	Random	$ICC(2,1)=\frac{\hat{\sigma}_T^2}{\hat{\sigma}_T^2+\hat{\sigma}_J^2+\hat{\sigma}_e^2}$
Subject by rater crossed design	Fixed	$ICC(3,1)=\frac{\hat{\sigma}_T^2}{\hat{\sigma}_T^2+\hat{\sigma}_e^2}$
Part B. Reliability of an average of *k* raters		
Nested: *n* subjects rated by *k* different raters	Random	$ICC(1,k)=\frac{\hat{\sigma}_T^2}{\hat{\sigma}_T^2+\hat{\sigma}_w^2/k}$
Subject by rater crossed design	Random	$ICC(2,k)=\frac{\hat{\sigma}_T^2}{\hat{\sigma}_T^2+(\hat{\sigma}_J^2+\hat{\sigma}_e^2)/k}$
Subject by rater crossed design	Fixed	$ICC(3,k)=\frac{\hat{\sigma}_T^2}{\hat{\sigma}_T^2+\hat{\sigma}_e^2/k}$

Table 12: Intraclass Correlation Coefficient (Shrout and Fleiss, 1979, cited in Lane et al., 2023)

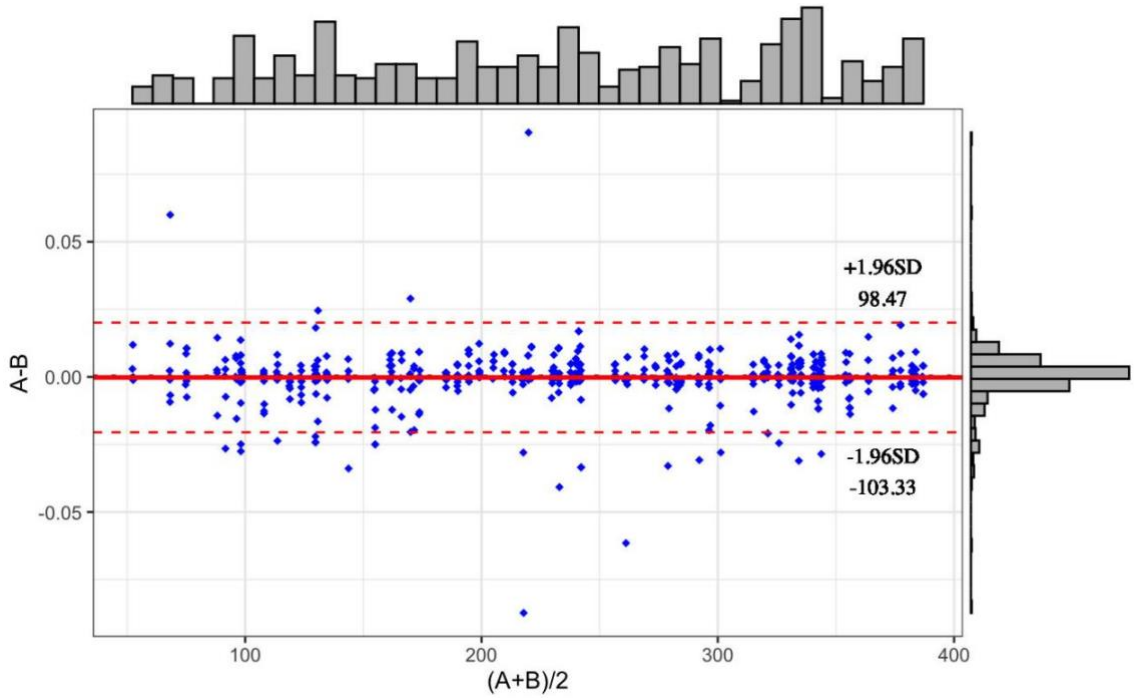

Figure 4: Bland-Altman plot for the current data set

6.3 Suggestions for annotation schemes

Here are some suggestions for improving the annotation schemes for landmarks that showed low inter-coder consistency in this analysis:

(1) *ev.b*

For annotating the landmark *ev.b*, it is recommended to enhance the upper bound of the spectrogram's view range to better visualize the higher frequencies. For an average adult female speaker, setting the upper bound at approximately 8000 Hz usually provides a clear view of the fricative. The landmark *ev.b* should be positioned where the energy becomes significantly dense, as illustrated in *Figure 5.1*. This can also be aided by identifying the starting point of the high-frequency vibrations in the waveform, as shown in *Figure 5.2*.

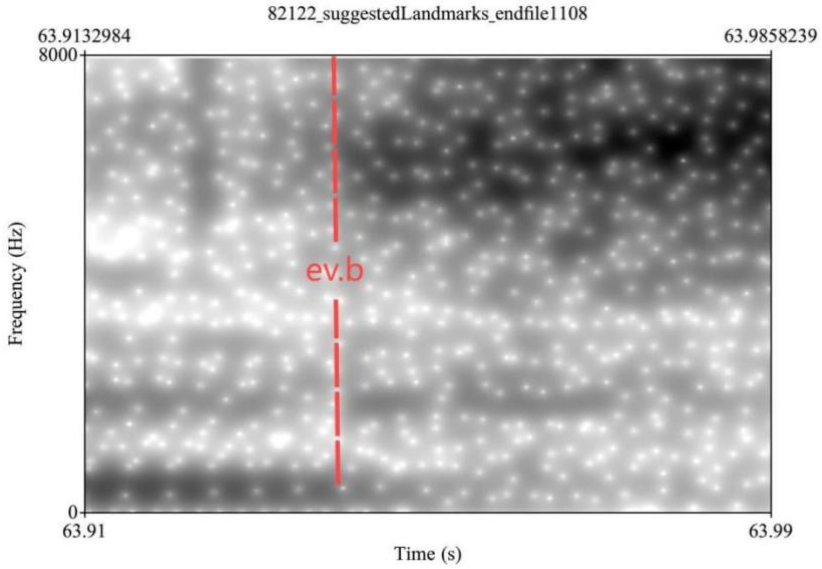

Figure 5.1: Spectrogram for landmark *ev.b*

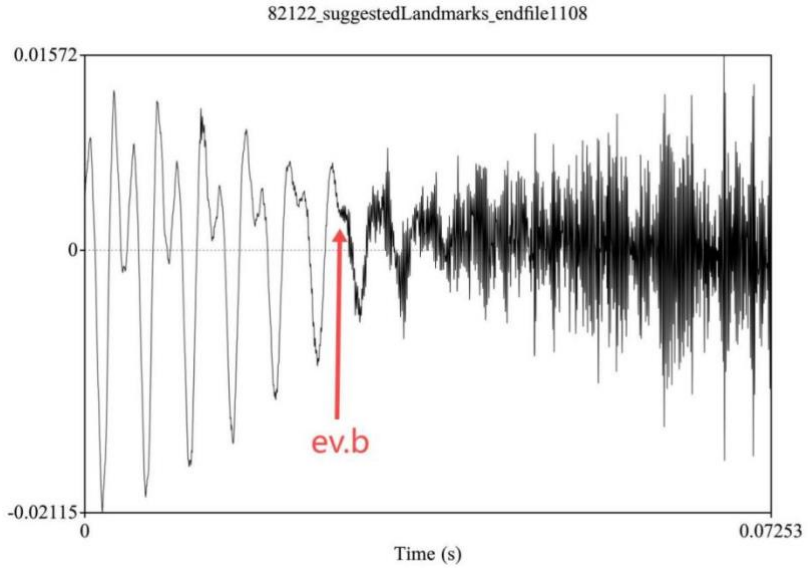

Figure 5.2: Waveform for landmark *ev.b*

(2) *ev.a*

Similarly, when annotating the landmark ev.a, it is also helpful to increase the upper bound of the spectrogram's view range. Although locating the endpoint of the suffix is more challenging than identifying the start point, it can still be determined by observing where the darkness in the higher frequencies ends (as shown in *Figure 6.1*). Additionally, the point where the amplitude of the high-frequency vibrations begins to decrease in the waveform can also serve as a useful reference (as shown in *Figure 6.2*).

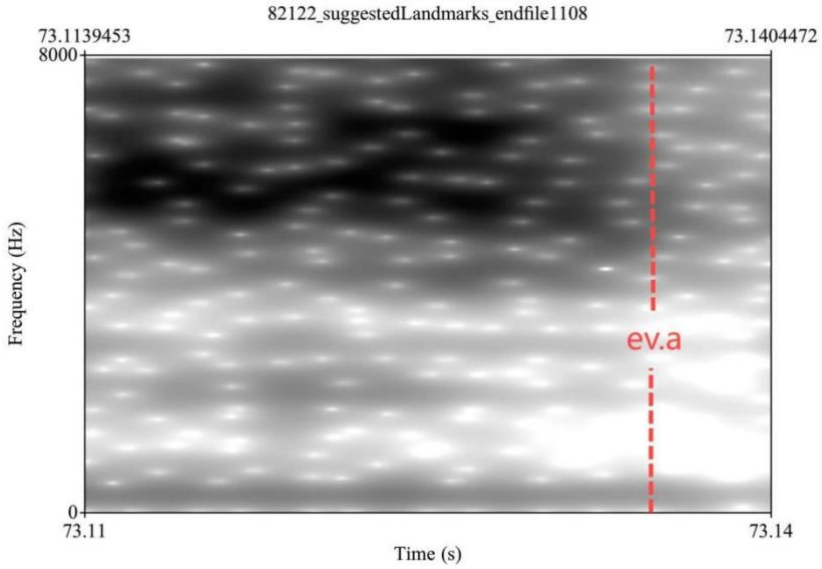

Figure 6.1: Spectrogram for landmark *ev.a*

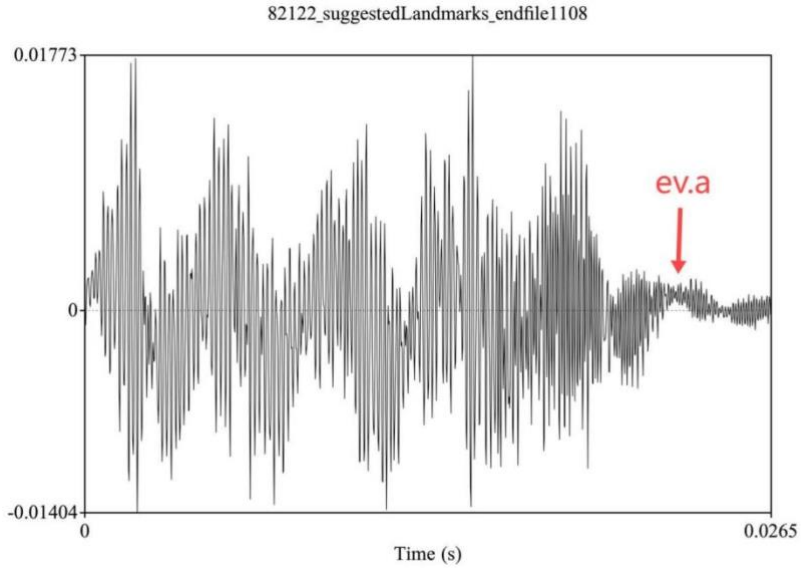

Figure 6.2: Waveform for landmark *ev.a*

(3) *eu.v*

For annotating the landmark *eu.v*, it is recommended to lower the upper bound of the spectrogram's view range to better visualize the second formant. For an average adult female speaker, setting the upper bound around 4000 Hz typically provides a clear view of the lower formants. The landmark *eu.v* should be placed at the endpoint of F2 (as shown in *Figure 7.1*). The end of the patterned periodicity of the vowel in the waveform can also serve as a potential reference (as shown in *Figure 7.2*).

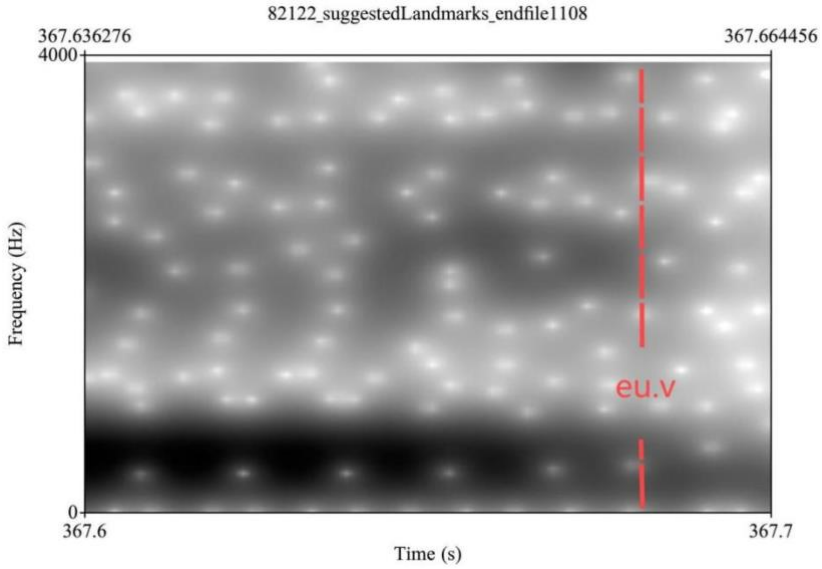

Figure 7.1: Spectrogram for landmark *eu.v*

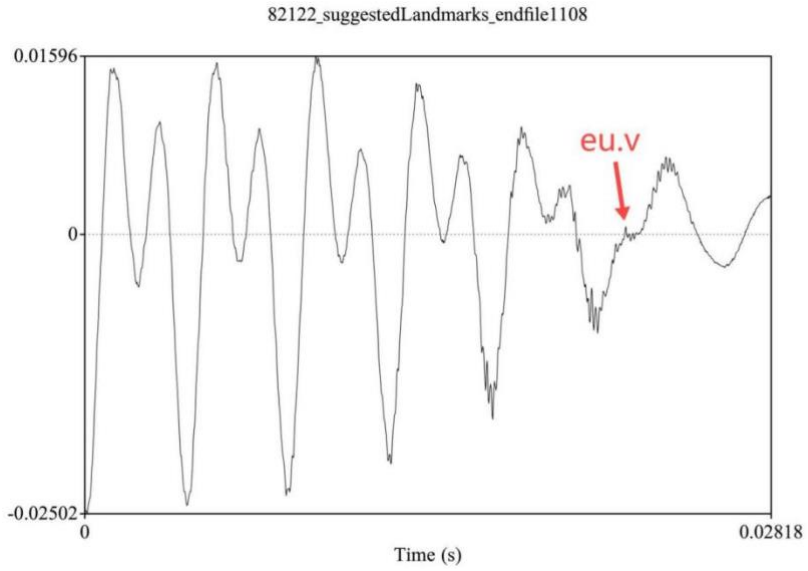

Figure 7.2: Waveform for landmark *eu.v*

(4) *sn*

The *sn* landmark might be the hardest to locate as it often blends with adjacent sounds. In some utterances, a shift in the lower formants may be visible (as shown in *Figure 8*), which can help identify where the initial consonant begins, especially when there is no preceding closure. If the visualization is less than ideal, adjusting the spectrogram blackness settings (i.e. *Maximum*, *Pre-*

emphasis and *Dynamic Compression*) in 'Advanced Spectrogram Settings' may enhance the visibility of the formants of intrest.

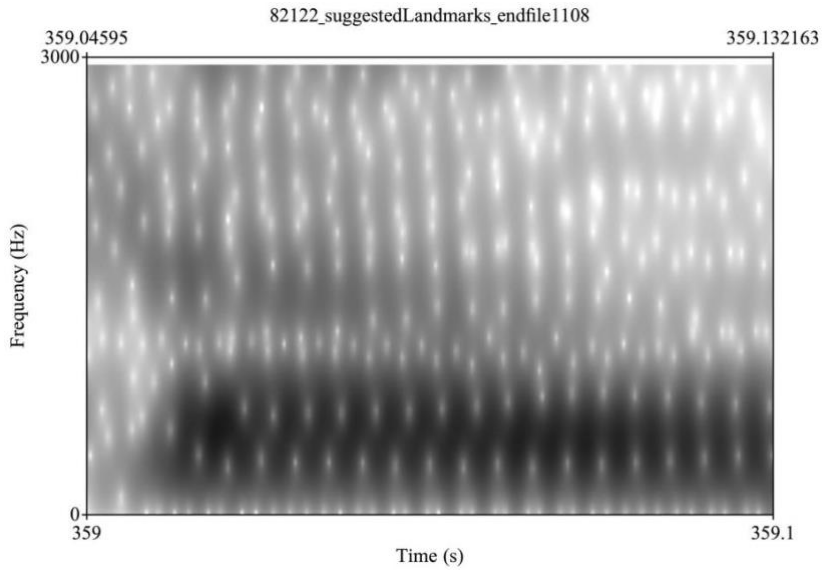

Figure 8: Spectrogram for landmark *sn*

If the *sn* landmark cannot be identified either through spectrogram analysis or by ear, it should be removed from the annotation. Given the consistently low inter-coder reliability for this landmark even after removing those deemed unlocatable, it may be worth considering excluding this landmark from future coding in the interest of efficiency.

Bibliography

Allen, G. & Hawkins, S. (1980). Phonological rhythm: definition and development. In G. Yeni-Komshian, J. Kavanagh & C. Ferguson (eds), *Child phonology, volume I: production*. New York: Academic Press. https://doi.org/10.1016/B978-0-12-770601-6.50017-6

Boersma, P., & Weenink, D. (2024). *Praat: doing phonetics by computer*. http://www.praat.org

Cai, R., Boersma, P., Yuen, I., Demuth, K., & Benders, T. (2024, July 2). Prosodic Clitics in English-speaking Children's Speech Production – An Acoustic Study. *Speech Prosody 2024*, Leiden, The Netherlands.

Carter, A., & Gerken, L. (2002). Do children's omissions leave traces?. *Journal of child language*, *31*(3), 561–586. https://doi.org/10.1017/s030500090400621x

Cohen, J. (1960). *A coefficient of agreement for nominal scales*. Educational and Psychological Measurement, 20(1), 37 – 46. https://doi.org/10.1177/001316446002000104

Coutanche, L. M. McMullen, A. T. Panter, D. Rindskopf, & K. J. Sher (Eds.), APA *handbook of research methods in psychology: Foundations, planning, measures, and psychometrics* (2nd ed., pp. 723–743). American Psychological Association.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297–334. https://doi.org/10.1007/BF02310555

Gerken, L. (1994). A metrical template account of children's weak syllable omissions from multisyllabic words. *Journal of Child Language, 21*, 565–84. https://doi.org/10.1017/S0305000900009466

Klein, H. (1981). Production strategies for the pronunciation of early polysyllabic lexical items. *Journal of Speech and hearing Research, 24*, 389–405. https://doi.org/10.1044/jshr.2403.389

Lane, S. P., Aslinger, E. N., & Shrout, P. E. (2023). Reliability. In H. Cooper, M. N.

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., Sonderegger, M. (2017) Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi. *Proc. Interspeech 2017*, 498-502. https://doi.org/10.21437/Interspeech.2017-1386

McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia medica, 22*(3), 276–282.

Rafieyan, V. (2016). Relationship between Acculturation Attitude and Translation of Culture-Bound Texts. *Journal of Studies in Education, 6*(2)

R Core Team (2013). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*(2), 420–428. https://doi.org/10.1037/0033-2909.86.2.420

Van Rossum, G., & Drake Jr, F. L. (1995). *Python reference manual*. Centrum voor Wiskunde en Informatica Amsterdam.

Wheeldon, L., & Lahiri, A. (1997). Prosodic Units in Speech Production. *Journal of Memory and Language, 37*(3), 356–381. https://doi.org/10.1006/jmla.1997.2517

Yuen, I., Demuth, K., & Shattuck-Hufnagel, S. (2022). Planning of prosodic clitics in Australian English. *Language, Cognition and Neuroscience, 37*(10), 1271–1276. https://doi.org/10.1080/23273798.2022.2060517

Appendix 1: Re-coding Manual

1. Overall

- **Overall goal:**
  Re-code approximately 20% of the utterances (tokens) of each annotated file.

- **Files:**
  For each participant, there is one 'longer' file (numbered as '82XX2', henceforward ***L-file***) and one 'shorter' file (numbered as '82XX1', henceforward ***S-file***). An *L-file* typically consists of 40pl+22sg=62 tokens; an *S-file* typically consists of 20sg+20pl=40 tokens.
  Hence, re-coding 20% of the tokens from an L-file and its corresponding S-file is by principle equivalent to re-coding 20% of the total tokens of that specific speaker.

- **"20%":**
  The number of tokens that should be re-coded for reliability analysis is determined by rounding 20% of the number of tokens in total to the nearest positive integer, as shown in *Table 1*:

Total number of tokens	**Number of tokens to be re-coded**
1-7	1
8-12	2
13-17	3
18-22	4
...	...

Table 1

- **Token selection:**
  Use a random number generator to determine which specific token of the group (for *S-files*)/sub-group (for *L-files*) should be re-coded according to *Table 2 and 3*.
  A specific token should not be "re-recoded" until all other tokens of the same series (e.g. 1-18 for an *S-file* or a1-a10 for an *L-file*) were re-coded for once in this round, so as to ensure that all tokens can be covered in the flow of the reliability check.
  An utterance should be skipped for the reliability check if it was labeled with disf/stut/notr, etc. In this case, generate another number to select a new token.

2. Re-coding: step-by-step

- **S-files**: 40*20% = **8 tokens** should be re-coded.

Step 1: List out the noun tokens and their landmarks. Practice items ('hat' and 'phone') should be excluded.
Step 2: Since all the noun tokens in S-files share the same landmarks, put the tokens into groups of with/without article (+art/-art).
Step 3: Calculate the number of tokens under each group.
Step 4: Determine the number of tokens that should be in reliability check from each sub-group according to *Table 1*.
Step 5: Select the tokens to be re-coded according to 1.4.

Tokens of a typical S-file (e.g. file 82051) is as shown in *Table 3*:

no.	verbs	noun tokens	landmarks	total number of tokens		number of tokens in reliability check	
				-art	+art	-art	+art
1	buy	ball	sn.c, sn.b1, snv	18	18	**4**	**4**
2	cover	bath					
3	cover	bean					
4	see	bib					
5	carry	bin					
6	see	bin					
7	carry	book					
8	see	boot					
9	carry	bowl					
10	buy	chair					
11	buy	cookie					
12	carry	cot					
13	cover	cup					
14	cover	cup					
15	buy	doll					
16	see	door					
17	buy	key					
18	cover	pen					

Table 3: S-file tokens

- **L-files:** 62*20% ≈ **12 tokens** should be re-coded.

Step 1: List out the noun tokens and their landmarks. Practice items ('hat' and 'phone') should be excluded.
Step 2: Put the noun tokens into groups according to their landmarks (a, b, c, d).
Step 3: Put the tokens into sub-groups of with/without article (+art/-art).
Step 4: Calculate the number of tokens under each sub-group.
Step 5: Determine the number of tokens that should be in reliability check from each sub-group according to *Table 1*.
Step 6: Select the tokens to be re-coded according to 1.4.

Tokens of a typical L-file (e.g. file 82052) is as shown in *Table 2*:

group	no.	verbs	noun tokens	landmarks	total number of tokens		number of tokens in reliability check	
					-art	+art	-art	+art
a	1	buy	light	sn, snv	10	10	**2**	**2**
	2	carry						
	3	cover	mango					
	4	see						
	5	buy	nappy					
	6	carry						
	7	cover	rock					
	8	see						
	9	cover	wombat					
	10	see						
b	1	buy	donut	sn.c, sn.b1, snv	8	10	**2**	**2**
	2	cover						
	3	buy	jacket					
	4	carry						
	5	cover	koala					
	6	see						
	7	buy	potty					
	8	carry						
	9	cover	table					
	10	see						
	11	carry	toy					
	12	see						

c	1	buy	balloon	su.c, su.b1, su.v, eu.v, sn, snv	8	--	**2**	0
	2	carry						
	3	buy	banana					
	4	carry						
	5	cover	giraffe					
	6	see						
	7	cover	tomato					
	8	see						
d	1	buy	computer	su.c, su.b1, su.v, eu.v, sn.c, sn.b1, snv	10	--	**2**	0
	2	carry						
	3	carry	guitar					
	4	see						
	5	cover	potato					
	6	see						
	7	buy	pyjamas					
	8	carry						
	9	buy	spaghetti					
	10	cover						

Table 2: L-file tokens

Appendix 2: Re-coding Log Data

File	Recoded tokens
82051	cover_cup_pl.E
	see_boot_pl.E
	buy_ball_pl.E
	buy_doll_pl.E
	buy_chair_sg.E
	buy_key_sg.E
	carry_cot_sg.E
	see_bin_sg.E
82052	carry_guitar_pl.E
	carry_balloon_pl. E
	cover_mango_pl.E
	carry_banana_pl.E
	see_rock_pl.E
	carry_toy_pl.E
	cover_koala_pl.E
	buy_pyjamas_pl.E
	see_mango_sg.E
	carry_potty_sg.E
	see_table_sg.E
	cover_wombat_sg.E
82071	cover_cup_sg.E
	cover_pen_sg.E
	carry_book_sg.E
	see_door_sg.E
	cover_bath_pl.E
	carry_bin_pl.E
	cover_bean_pl.E
	carry_bowl_pl.E
82072	cover_donut_pl.E
	carry_computer_pl.E
	see_wombat_pl.E
	cover_tomato_pl.E
	see_potato_pl.E
	see_giraffe_pl.E
	cover_rock_pl.E
	buy_jacket_pl.E
	buy_nappy_sg.E
	buy_potty_sg.E
	buy_light_sg.E
	cover_table_sg.E
82091	buy_doll_pl.E
	buy_ball_pl.E
	carry_book_pl.E
	see_bib_pl.E
	buy_chair_sg.E
	cover_cup_sg.E
	buy_key_sg.E
	cover_cup_sg.E
82092	carry_light_pl.E
	see_toy_pl.E
	see_mango_pl.E
	carry_pyjamas_pl.E
	buy_banana_pl.E
	cover_giraffe_pl.E
	carry_jacket_pl.E
	buy_donut_sg.E
	cover_rock_sg.E
	carry_potty_sg.E
	cover_wombat_sg.E
82111	see_boot_sg.E
	see_bib_sg.E
	see_door_sg.E
	cover_pen_sg.E
	buy_bikkie_pl.E
	cover_bath_pl.E
	see_bin_pl.E
	carry_cot_pl.E
82112	cover_spaghetti_pl.E
	buy_jacket_pl.E
	carry_nappy_pl.E
	buy_light_pl.E
	cover_potato_pl.E
	see_tomato_pl.E
	buy_balloon_pl.E
	see_koala_pl.E
	cover_table_sg.E
	carry_light_sg.E
	see_table_sg.E
	see_wombat_sg.E

Appendix 3: Scripts - where to find them and what do they do

The Python and R scripts as well as relevant files can be found in:

https://amsuni-my.sharepoint.com/:f:/g/personal/yuying_zhu_student_uva_nl/EnoVTXpF-Z5Lu6AdfGq_hSgB7gCQmuYw3wj2uPTQQ2Itvw?e=crSBQg

R script:

The R Markdown file named "Reliability_R.Rmd" can be used for calculation of descriptive statistics and data visualization. The R Markdown file must be stored in the same folder with other data files to function properly.

In addition, some trial codes for ICC, Cronbach's alpha and Bland-Altman plots are included.

Python scripts:

The *Python* file named "reliability1.1.py" requires a series of files as input that should be stored in the same folder as the script(s):

1. A set of raw data files in csv form (see *Appendix 4*)
2. A csv file named "listfilenumber.csv"
3. A csv file named "nountarget.csv"
4. A csv file named "listtypes.csv"

In principle, "reliability1.1.py" can automatically process any similar raw data in CSV format as long as:

1. The corresponding csv files of the two coders' annotations on the same sound file are named correctly. For instance, for sound file 82051, the two csv files should be named as "82051_suggestedLandmarks.csv" and "82051_suggestedLandmarks_recoded.csv".

2. The names of the files to be processed are logged in "listfilenumber.csv". For instance, for file 82051, log "82051" into "listfilenumber.csv".

3. There are no commas in the annotation texts (this can be easily fulfilled with the *Find and Replace* function in excel).

After running the script, the processed data can be found in a file in the same folder named “reliability output all_T.csv” (as shown in *Figure Appendix 3*). The features involved in the analysis can be added or deleted in “nountarget.csv” and “listtypes.csv”, with the codes also be changed accordingly. Another csv file named “kappa repeat.csv” stores the repeatedly calculated (set at 50 times) kappa values of each category of interest, the second last row shows the average of the 50 kappa values, and the last row shows the variance of each group.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	file number	target	landmark type	landmark	absolute time difference	noun	landmark set	prosodic structure	initial consonant of stressed syllable	singular/plural	consistency	time coder1	time coder2	kappa index 1	kappa index 2
2	82111	see_bib_sg.E	landmarks_tssv		0.000228	bib	b	s	plosive	singular	consistent	194.3657	194.365928	0	0
3	82111	see_bib_sg.E	landmarks_tev.b		0.000551	bib	b	s	plosive	singular	consistent	194.521126	194.520575	0	0
4	82111	see_bib_sg.E	landmarks_tev.a		0.004104	bib	b	s	plosive	singular	consistent	194.575731	194.571627	0	0
5	82111	see_bib_sg.E	landmarks_tsu.v		0.001005	bib	b	s	plosive	singular	consistent	194.584611	194.585616	0	0
6	82111	see_bib_sg.E	landmarks_tsn.c		0.008485	bib	b	s	plosive	singular	inconsistent	194.628006	194.619521	0	1
7	82111	see_bib_sg.E	landmarks_teu.v		0.008485	bib	b	s	plosive	singular	inconsistent	194.628006	194.619521	0	1
8	82111	see_bib_sg.E	landmarks_tsn.bl		0.001782	bib	b	s	plosive	singular	consistent	194.704463	194.706245	0	0
9	82111	see_bib_sg.E	landmarks_tsnv		0.000158	bib	b	s	plosive	singular	consistent	194.730637	194.730479	1	1
10	82051	see_bin_sg.E	landmarks_tssv		0.002767	bin	b	s	plosive	singular	consistent	383.594665	383.597432	0	0
11	82051	see_bin_sg.E	landmarks_tev.b		0.005013	bin	b	s	plosive	singular	inconsistent	383.704371	383.709384	1	0
12	82051	see_bin_sg.E	landmarks_tev.a		0.009243	bin	b	s	plosive	singular	inconsistent	383.762216	383.752973	1	0
13	82051	see_bin_sg.E	landmarks_tsu.v		3.30E-05	bin	b	s	plosive	singular	consistent	383.763482	383.763515	1	1
14	82051	see_bin_sg.E	landmarks_tsn.c		0.002754	bin	b	s	plosive	singular	consistent	383.810389	383.807635	0	0
15	82051	see_bin_sg.E	landmarks_teu.v		0.002754	bin	b	s	plosive	singular	consistent	383.810389	383.807635	1	1
16	82051	see_bin_sg.E	landmarks_tsn.bl		0.001906	bin	b	s	plosive	singular	consistent	383.88788	383.889786	1	1
17	82051	see_bin_sg.E	landmarks_tsnv		0.000521	bin	b	s	plosive	singular	consistent	383.90701	383.906489	1	1
18	82071	carry_book_sg.E	landmarks_tssv		0.003386	book	b	s	plosive	singular	consistent	118.646149	118.649535	0	0
19	82071	carry_book_sg.E	landmarks_tev.b		0.001523	book	b	s	plosive	singular	consistent	118.850488	118.852011	1	1
20	82071	carry_book_sg.E	landmarks_tev.a		0.004292	book	b	s	plosive	singular	consistent	118.912108	118.9164	1	1
21	82071	carry_book_sg.E	landmarks_tsu.v		0.008203	book	b	s	plosive	singular	inconsistent	118.915309	118.923512	0	1
22	82071	carry_book_sg.E	landmarks_tsn.c		0.001793	book	b	s	plosive	singular	consistent	118.959323	118.961116	1	1
23	82071	carry_book_sg.E	landmarks_teu.v		0.001793	book	b	s	plosive	singular	consistent	118.959323	118.961116	1	1
24	82071	carry_book_sg.E	landmarks_tsn.bl		0.0005	book	b	s	plosive	singular	consistent	119.031347	119.031847	1	1
25	82071	carry_book_sg.E	landmarks_tsnv		0.000372	book	b	s	plosive	singular	consistent	119.068159	119.067787	1	1
26	82111	see_boot_sg.E	landmarks_tssv		0.001551	boot	b	s	plosive	singular	consistent	184.816077	184.814526	1	1
27	82111	see_boot_sg.E	landmarks_tev.b		0.003911	boot	b	s	plosive	singular	consistent	184.985559	184.98947	1	1
28	82111	see_boot_sg.E	landmarks_tev.a		0.000621	boot	b	s	plosive	singular	consistent	185.034046	185.033425	1	1
29	82111	see_boot_sg.E	landmarks_tsu.v		0.001163	boot	b	s	plosive	singular	consistent	185.043825	185.044988	1	1
30	82111	see_boot_sg.E	landmarks_tsn.c		0.001559	boot	b	s	plosive	singular	consistent	185.0904	185.088841	0	0
31	82111	see_boot_sg.E	landmarks_teu.v		0.001559	boot	b	s	plosive	singular	consistent	185.0904	185.088841	1	1
32	82111	see_boot_sg.E	landmarks_tsn.bl		0.000643	boot	b	s	plosive	singular	consistent	185.179674	185.180317	0	0
33	82111	see_boot_sg.E	landmarks_tsnv		0.000558	boot	b	s	plosive	singular	consistent	185.201718	185.20116	1	1

Figure Appendix 3: ‘reliability output all_T.csv’

The files named “duration_usv.py” and “duration_3sg.py” prints the kappa results of unstressed vowels and 3rd person singular suffixes as output respectively. Since both scripts take “reliability output all_T.csv” as input, they must be run after the process of “reliability1.1.py” was ended.

Appendix 4: Files

All files involved in the reliability analysis can be found in:
https://amsuni-my.sharepoint.com/:f:/g/personal/yuying_zhu_student_uva_nl/Eoj-vBY9G99BmDtcS8Ni5PAB0wNlfTfolNy92UymHnd9fA?e=aGV1QV

An overview of the consistency between coders can be found in:
https://amsuni-my.sharepoint.com/:x:/g/personal/yuying_zhu_student_uva_nl/EWunkcgVtE5ApV9hp32zsXoBQEubq1wlvhsDT9THQt6Gfg?e=eCYvMM

An overview of the processed data can be found in:
https://amsuni-my.sharepoint.com/:x:/g/personal/yuying_zhu_student_uva_nl/Ec7d2mtEqKlDtsMw3jZ1LW0B4uu7yxIw2YKk7OGdEmu3UQ?e=NWnytO

The Python and R scripts as well as relevant files can be found in:
https://amsuni-my.sharepoint.com/:f:/g/personal/yuying_zhu_student_uva_nl/EnoVTXpF-Z5Lu6AdfGq_hSgB7gCQmuYw3wj2uPTQQ2Itvw?e=crSBQg

The raw data of the reliability analysis can be found in:
CSV: https://amsuni-my.sharepoint.com/:f:/g/personal/yuying_zhu_student_uva_nl/ElimrpRtBThDkhGXjKhayKUB_V7T0VaeUPeDTAxOyqMK6Q?e=a91nSy
TextGrid: https://amsuni-my.sharepoint.com/:f:/g/personal/yuying_zhu_student_uva_nl/EjHHnr4dVtpDuOSEYtoPtBwBfHGQwoUN6aS8kvJKNzff_g?e=gWMXgm

Appendix 5: Interpretation Criteria for Cohen's kappa

Value of Kappa	Level of Agreement	% of Data that are Reliable
0 – 0.20	**None**	0–4%
0.21 – 0.39	**Minimal**	4–15%
0.40 – 0.59	**Weak**	15–35%
0.60 – 0.79	**Moderate**	35–63%
0.80 – 0.90	**Strong**	64–81%
Above 0.90	**Almost Perfect**	82–100%

Interpretation of Cohen's kappa (McHugh, 2012)

Value of Kappa	Level of Agreement
< 0	**Poor Agreement**
0 – 0.20	**Slight Agreement**
0.21 – 0.40	**Fair Agreement**
0.41 – 0.60	**Moderate Agreement**
0.61 – 0.80	**Substantial Agreement**
0.81 – 1	**Almost Perfect Agreement**

Interpretation of Cohen's kappa (Rafieyan, 2016)