ANNOTATION OF DISFLUENCIES IN CHILD SPEECH

Valentin Kany[1]*, Jürgen Trouvain*[2]

Language Science and Technology, Saarland University, Saarbrücken, Germany
[1]*valentin.kany@uni-saarland.de,* [2]*trouvain@lst.uni-saarland.de*

Abstract: This study introduces a scheme for annotating disfluencies in the speech of preschool children. Based on this scheme, an individual fluency profile is compiled and presented that provides an overview of various fluency-related aspects of the child's speech. Such a profile is intended to be used and integrated into language proficiency assessments. The study reports preliminary results of the fluency behaviour of a small sample of children (n=10) and demonstrates how the speech fluency profile can be put into use. A high degree of individuality in the usage pattern of different disfluency types is found, which underlines the importance of the individual view the fluency profile offers. Thus, the behaviour of two children with respect to their fluency is compared. Two radar charts derived from their fluency profile highlight the individual differences. The study discusses the meaning of this fluency assessment in the field and proposes a possible approach to integrate it into general language proficiency assessments.

1 Introduction

Language Proficiency Assessment (LPA) for preschool children is ubiquitous in Germany and will become mandatory in more and more federal states over the upcoming years [1]. Up until now, LPA is done manually, with pen and paper, which is highly subjective, inconsistent, and creates a test setting for the children that hinders natural speech production. To tackle this problem, a serious-game-based test method was developed, in which the children talk with a virtual character "Wuschel" [2]. With this test method, the children's speech is analysed using various measures of vocabulary and grammar. Beyond that, speech fluency was found to correlate with speech competence [3]. Thus, it could serve as a further factor in the LPA test paradigm for preschool children that is yet to be investigated.

To analyse the fluency of children, we propose an annotation method suitable for the game-based test method used here. It covers a variety of phenomena and indicators of fluency and disfluency in the children's speech produced during the game. Using this annotation scheme, we develop individual speech fluency profiles for each child. These profiles serve as an overview of the child's performance in terms of speech fluency. They highlight strengths and weaknesses at a glance and reveal individual patterns. This is particularly important to get a clear picture of the child's abilities to address their deficits and support them with an individual plan, which is the main goal of LPAs.

2 Data

The raw data is collected in German daycare centres where children (age: 4.5-6 years) are recorded as part of an LPA while playing the Wuschel game [2]. The recording sessions take place in a separate room (e.g. the daycare centre's staff meeting room) to minimise the background noise produced by other children. During the session, at least one of our project's team

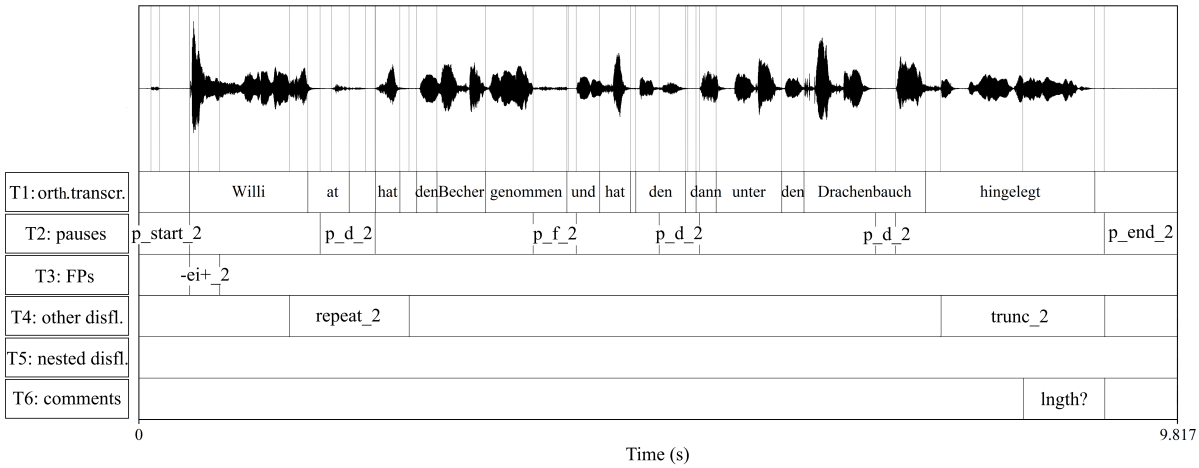

Figure 1 – Example utterance with annotations on 6 tiers. See text for details.

members and one of the daycare centre's staff, who functions as a confidant to the child, are present in the room. Participants are recorded by the built-in microphone of the iPad (9th gen) used to play the game. This way, no microphone is visible and the children do not feel uncomfortable as they do not know they are being recorded. The age range of the participants covers the span in which children begin to use filler particles such as "äh" or "ähm" in an adult-like manner [4]. The recordings of 10 children were analysed in this study, 5 of them grow up with German as their native language (L1) and 5 with German as their second language (L2). The game consists of a coherent story with 28 game scenarios. In every scenario, the child has to answer two questions in order for the game to progress. The second question is always formulated as a follow-up question to the first to give the child the opportunity to further elaborate on their answer. Thus, every child produced a total of 56 recorded segments per session. An average playthrough of the game lasts for a total of about 30 minutes but yields only about 8 minutes of recorded segments. These still contain speech from the adults supporting the child as well as some background noises and interaction pauses. We cleaned the data by muting non-child speech, which results in an average of 3:02 minutes of articulation time per child (pauses excluded).

3 Methodology

3.1 Annotation

The proposed annotation scheme is based on [5] and uses the TextGrid function of Praat [6] (Figure 1). The TextGrid contains a total of 6 tiers. The manually generated orthographic transcript on tier 1 was aligned using WebMAUS [7] and left uncorrected, as it just serves as a means of orientation. The rest of the tiers (T2-T6) were annotated manually.

Tier 2 is used for pause and inter-pause interval annotations. Here, the annotators mark the articulation phases of both the child and external speakers present during the recording session (e.g. the experimenter or daycare centre staff) along with the pauses. The annotation of external speakers' articulation phases and pauses helps to further analyse the interaction between the child and external speakers and provides useful information on the degree of support a child needed during the LPA. For the annotation of the (silent) pauses, we followed the basic approach of [5], who defined silent pauses as phases with either no audible acoustic noise or non-talking phases in which respiratory or articulatory activities can still be found. Filler particles (FPs) were not classified as pauses but received attention on a separate tier in T3. The pause annotations within talking turns are extended by the annotator's judgment on whether

they perceived the pause to be disfluent (d) or part of a fluent phase (f). Additionally, 4 types of pauses between turns are considered separately:

- Pauses at the beginning of the segment, before the child starts their answer
- Pauses at the end of the segment, after the child finished their answer
- Pauses during turn taking between the child and external speakers (before the child starts to talk and before an external speaker starts to talk)

The duration of pauses before and after finished utterances is influenced by many factors, which are not directly linked to the child's fluency. These factors include the process of understanding the task, courage to give an answer, and the social bond between the interlocutors [8]. Thus, these types of pauses received no further judgment with respect to their effect on the perceived fluency and are excluded from the current analysis.

Tier 3 was used for the annotation of FPs. The annotators differentiate between four types: particles consisting of only a vowel ("äh"), a vowel and a nasal ("ähm"), just a nasal consonant ("hm"), and a discourse particle consisting of a diphthong ("ei"). Further, information on the speech context of these FPs is annotated. If the particle is preceded by speech, "+" is annotated in front of the particle type. If it is preceded by silence, "-" is used respectively. The same check is done for the context following the particle, and either "+" or "-" is attached to the annotation of the particle type (i.e. a vowel-only filler particle that is preceded by silence and followed by speech is annotated as "-äh+").

Other disfluencies are annotated on two tiers (T4 and T5), with T5 being used to cover possible nestings. The disfluencies considered are repairs (speech errors which are corrected shortly afterwards), truncations (abandonments of syllables, words or clauses at some point during the utterance), lengthenings (prolongations of speech sounds), and repetitions (reiterations of words). Lastly, tier 6 serves as an option for additional comments. All labels concerning the flow of speech (T2-T5) are extended by the current prompt number (1 or 2), as this might have an effect on the fluency of the child.

3.2 Fluency Profiles

Since LPA is about carrying out individual assessments and providing an individually tailored plan for the child, having an overview of the overall performance of a child is indispensable. Also, the general results from Table 1 suggest individual disfluency patterns. An individual in-depth profile of the child offers a neat way to determine the children's level of fluency and behaviour in case of disfluency. This fluency profile is based on the annotation of the child's speech production, summarised in a table-like overview (Figure 2) and visualised in radar charts (Figure 3).

A sample fluency profile of one child is shown in Figure 2. The fluency profile starts with biographic data of the child at the top. It includes information like age, contact time with German, and type of language acquisition (L1 or L2). All factors are important to consider in the analysis of the child, as they might have an influence on the child's fluency (e.g. articulation rate grows with age [9]). Thus, researchers usually want to compare an individual child to other children with the same prerequisites.

Below the biographic data, we included all fluency-related measurements [10] that are derivable from our annotations. Additionally, we used the Praat Scripts from [3] to be able to determine speaking and articulation rate, as [11] demonstrated their usability for German child speech. To improve the readability of the profile, we divided all measurements into different categories (C1-C5). We added the mean values of all children in our database to check

Biographic Data		Test run	
Child 1212bb63-c357-4500-8423-6fa4ef44c9b3		Date of recording	01/02/2024
Age	5;5	Daycare centre	XXX
German: L1 or L2	L2	Daycare centre type	no focus
Contact time with German	2;5	Run number	1

C1 Speech duration		Mean (all children)	
Articulation time	04:41 min	02:48 min	↗
External speech time	03:01 min	02:03 min	–
Speaking rate	2.69 syl/s	2.77 syl/s	–
Articulation rate	3.19 syl/s	3.09 syl/s	–
Longest articulation phase	9.19 s	6.00 s	↗
Mean articulation phase	1.68 s	1.43 s	–

C2 Pauses	total	per minute	Mean (all children) total	Mean (all children) per minute	
Number of all pauses	91	19.46	52.2	15.27	–
Number of fluent pauses	44	9.41	18.5	5.07	–
Number of disfluent pauses	47	10.05	33.7	10.20	–
Total pause duration	51.58 s		23.49 s		↘
Ratio pause duration:articulation time	0.18		0.12		–

C3 Filler Particles	total	per minute	Mean (all children) total	Mean (all children) per minute	
Number of all filler particles (FP)	29	6.20	16.1	5.89	–
Number of "äh"	1	0.21	5.5	2.08	–
Number of "ähm"	3	0.64	2.3	0.65	–
Number of "hm"	9	1.92	2.5	1.06	–
Number of "ei"	16	3.42	2	0.51	↘
Total filler particle duration	15.04 s		10.87 s		–
Ratio FP duration:articulation time	0.05		0.07		–

C4 Other disfluencies	total	per minute	Mean (all children) total	Mean (all children) per minute	
Number of all other disfluencies	59	12.62	21.1	6.23	↘
Number of repairs	9	1.92	3.9	1.27	–
Number of truncations	22	4.70	6.1	1.80	↘
Number of lengthenings	12	2.57	5.7	1.62	–
Number of repetitions	16	3.42	5.4	1.54	–
Total other disfluency duration	94.04 s		29.24 s		↘
Ratio oth. disfl. dur.:articulation time	0.34		0.14		↘

Figure 2 – Sample fluency profile of Child 1, as discussed in section 4.2.

if the child performed better or worse than the average values across all children in the sample. A thin-lined, upward pointing arrow indicates that the child performed at least 1.5 times the standard deviation above the mean, a thick-lined, downward pointing arrow shows that the child performed at least 1.5 times the standard deviation below the mean. This helps to quickly find areas in which the child still might need some support to improve and catch up with other children (currently only n = 10). If the database was larger, the comparison group would have to be limited to fit the child's biographic data.

The first category (C1) contains general duration-related information of the child's recording. The total articulation time and external speech time provides insight into the talkativity of the child and whether they needed much support by external speakers during the recording session. This information is crucial, because in general, a more talkative child has more opportunities to become disfluent than a timid child that hardly produces any coherent sentences at all. Similar to that, the values of the longest and mean duration of an articulation phase indicates if a child is able to produce longer coherent utterances or not. As final measurements in this category, we provide the speaking and articulation rate of the child. According to [12], these are among the most suitable measures to determine prosodic competence in a language.

The second category (C2) includes all results related to the pause annotations. Especially the number of disfluent pauses per second of articulation time allows for a judgment on the child's fluency, as the disfluent pauses have already been judged to interrupt the child's fluency by the annotators beforehand.

In the third category (C3), the child's behaviour with respect to the usage of FPs is depicted. While the overall frequency of FPs could be a relevant indicator for the child's fluency, the frequency of the filler particle types is rather useful to find individual usage patterns or give an indication of whether or not the child has already acquired the usage of a certain type.

The last category (C4) covers all the other disfluencies that were annotated on tier 4 and 5. Similar to the FPs in C3, the total number of other disfluencies per second can contribute to the overall impression of the fluency of the child, whereas the individual disfluency types rather reveal specific usage patterns and preferred ways of hesitation in the child's speech. The fluency profile is closed off by two radar charts (Figure 3), which provide a way to directly compare the individual child to the mean and see how they differ in their (dis)fluency patterns from others.

4 Results

4.1 General Results

The data shows that all analysed children produce FPs. However, there is a large standard deviation in the amount of FPs used per child, which leads to the impression that children either produce them frequently or barely at all (Table 1c). The high variation can also be found across other types of disfluencies (Table 1d). The production of the various types of disfluencies seems to be a highly individual aspect, which makes the analysis of individual speech fluency profiles more interesting. The standard deviations in Table 1d suggest that the timidity of some children still plays a role, even in our gamified LPA. Some children produce less speech and need more encouragement by external speakers. The contexts in which the different types of FPs were produced show that the FPs are not produced arbitrarily by the children (Table 1c). Some FPs have a high affinity with certain pause/speech contexts, e.g. 15 of 20 occurrences of the discourse particle "ei" start after a pause just before speech (-ei+). This finding suggests that children have internalised the specific function of some FPs. However, their clear tendency towards the production of one single FP type shows that they have not acquired the full repertoire of FPs yet. The prompt number seems to play a minor role, as there are no significant differences in

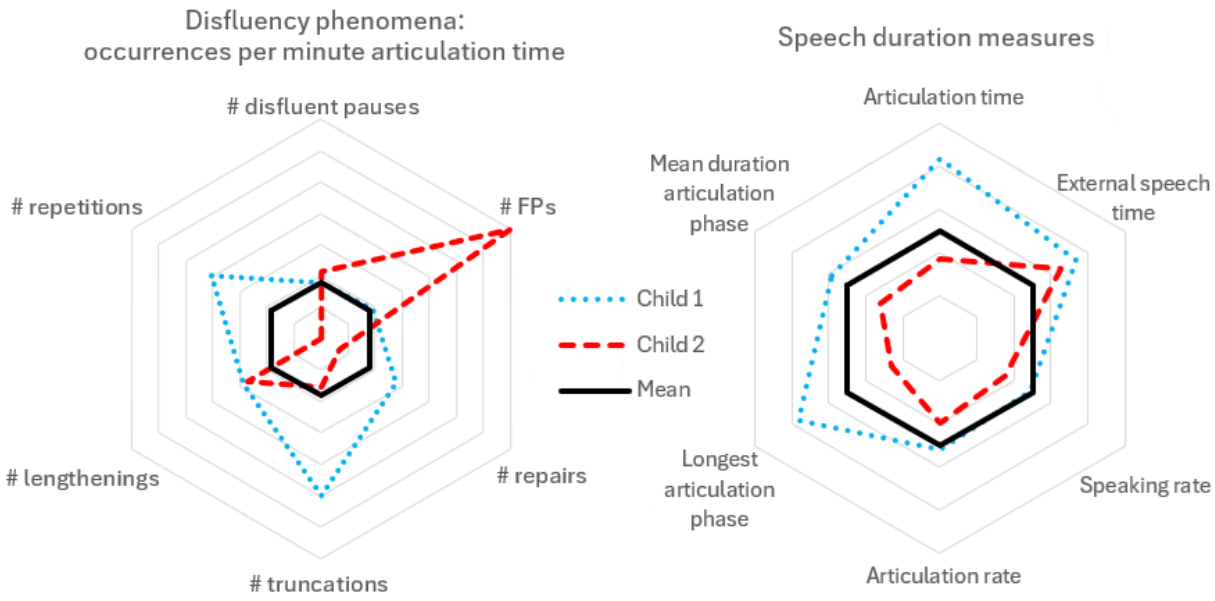

Figure 3 – Radar charts for Child 1 and Child 2 as they are attached to a single child's fluency profile to see the child's behaviour with regards to fluency-related measures. Values are normalised by the measures' respective means. This way, all measures can be compared on one scale and individual patterns in relation to the mean are revealed. Absolute values can be taken from the fluency profile itself.

1a	**sum dur. segm.**	**dur. child speech**	**dur. external sp.**	**dur. pauses**	
	7:54 min (1:49)	3:02 min (0:57)	1:43 min (0:54)	3:09 min (0:56)	
1b	**no. pauses**	**no. fl. pauses**	**no. disfl. pauses**	**dur. fl. pauses**	**dur. disfl. pauses**
	52.2 (42.02)	18.5 (18.94)	33.7 (24.44)	232 ms (94)	593 ms (268)
1c	**no. FPs**	**no. äh** **[++, --, +-, -+]**	**no. ähm** **[++, --, +-, -+]**	**no. ei** **[++, --, +-, -+]**	**no. hm** **[++, --, +-, -+]**
	12.3 (15.34)	5.5 (8.53) [0.1, 2.9, 0.8, 1.7]	2.3 (4.62) [0.2, 1.1, 0.4, 0.6]	2.0 (4.99) [0.0, 0.5, 0.0, 1.5]	2.5 (5.25) [0.0, 2.3, 0.0, 0.2]
1d	**no. other disfl.**	**no. repairs**	**no. truncations**	**no. lengthenings**	**no. repetitions**
	21.1 (17.98)	3.9 (3.18)	6.1 (6.57)	5.7 (7.85)	5.4 (5.62)

Table 1 – All values are mean values **per child**, standard deviation in parentheses. **a:** Distribution of the duration of the 56 recorded segments on child speech, external speech, and speech pauses. **b:** Results of the analysis on speech pauses of the annotated children. **c:** Results of the analysis on FPs of the annotated children. For all particles, their mean occurence per child in each analysed speech/pause context is given in square brackets. **d:** Results of the analysis on other disfluencies.

the disfluency patterns found in both prompts. This could be due to the fact that the novelty of a scene and its challenge in the first prompt is balanced out by the higher elaborateness the child is asked for in their answer to the second prompt.

4.2 Individual Fluency Profiles Results: Comparison of two children

We compare the profile of two children with similar prerequisites (both with German as L2, both 5.5 years old and around 2 years of contact time) who behaved very different from each other. Child 1 was very talkative and produced more than twice the amount of speech of Child 2 (4:41 minutes vs. 2:05 minutes). The question arises how this affects the fluency profile of both children and if the fluency profile of Child 2 still allows for any conclusion regarding their fluency, despite of the short articulation time. Both articulation rate and speaking rate of Child 2 are lower than for Child 1. The speaking rate lies almost 1 syllable per second below the mean, which is an indicator of disfluent speech, independent of the amount of speech material produced by the child. Also, the value on "Longest articulation phase" for Child 1 (9.19 seconds)

lies clearly above average, for Child 2 clearly below average (3.11 seconds). This measure is, in contrast to most of the other measures, focused on the child's direct ability to produce coherent, fluent speech over a period of time, rather than analysing the child's disfluencies and deducting their fluency from that. Thus, this measure is not boosted by the circumstance of the child being not very talkative. Other measurements from the categories C2-C4 are linked to the amount of produced speech material, but were normalised with respect to the total articulation time of the child. Regarding the number of disfluent pauses per second, both children performed similarly (0.17 pauses per second vs. 0.20 pauses per second). While Child 1 used fewer FPs than Child 2 (0.10 vs. 0.38), Child 1 produced more other disfluencies than Child 1 (0.21 vs. 0.07). Overall, the radar charts (Figure 3) support the impression that Child 1 speaks more fluently than Child 2, independent of the fact that Child 2 produced less speech material and thus had fewer opportunities to become disfluent.

5 Discussion and Conclusion

The analysed data shows that the usage pattern of the different types of disfluencies in child speech strongly depends on the individual speaker. This underlines the necessity of investigating children's disfluency patterns on an individual level. Also, some children are less talkative than others and therefore have less room for the production of disfluencies. This should be kept in mind when assessing a child's speech fluency. Both aspects are addressed by the fluency profile introduced here. The comparison of the two children shows that fluency assessments can be made, regardless of the difference in articulation time between the children. Still, it is unclear to which degree the simple normalisation by division by total articulation time can compensate for the bias introduced by timid children in general. Apart from that, there are some further issues that need to be addressed in future work. While the profiles allow for a first glimpse at the child's abilities and fluency, it cannot immediately be seen which factor contributes in which way to the perceived fluency. Therefore, perception experiments have to be conducted to gain insights into the effects the different measures from this study have on speech fluency. Participants could rate the children based on their perceived fluency. The ratings should address the categories included in the fluency profiles introduced in this study. This way, perceived fluency profiles of the children can be created. With this knowledge, weights could be added to the measures in the annotation-based fluency profile introduced here. Eventually, an overall fluency score could then be derived from the weighted fluency profile, which would be key to integrating the fluency assessment into general LPA.

Acknowledgements:

We are grateful to Julia Schu for annotating our data and Diana Davidson for cleaning the data. We also would like to thank Bernd Möbius for his valuable comments during the writing process of this paper.

References

[1] LISKER, A.: *Sprachstandsfeststellung und Sprachförderung im Kindergarten sowie beim Übergang in die Schule. Expertise im Auftrag des Deutschen Jugendinstituts*, 2010. URL http://www.dji.de/bibs/Expertise_Sprachstandserhebung_Lisker_2010.pdf.

[2] ROCHE, J., S. HABERZETTL, G. PAGONIS, M. JESSEN, and N. WEIDINGER: *Serious Games in der Sprachstandsermittlung*, pp. 340–358. Narr Francke Attempto Verlag, 2019. doi:http://dx.doi.org/10.22028/D291-35846.

[3] DE JONG, N. H., J. PACILLY, and W. HEEREN: *Praat scripts to measure speed fluency and breakdown fluency in speech automatically. Assessment in Education: Principles, Policy & Practice*, 28(4), pp. 456–476, 2021. doi:10.1080/0969594X.2021.1951162. URL https://doi.org/10.1080/0969594X.2021.1951162. https://doi.org/10.1080/0969594X.2021.1951162.

[4] VAN DER WEGE, M. M. and E. C. RAGATZ: *Learning to be fluently disfluent*. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, vol. 26, p. 1647. 2004.

[5] MUHLACK, B.: *Filler particles: phonetic details, cross-linguistic comparisons, and the recall effect*. Ph.D. thesis, Saarland University, 2023. doi:http://dx.doi.org/10.22028/D291-42474.

[6] BOERSMA, P. and D. WEENINK: *Praat: doing phonetics by computer (version 6.4.04)*. 2024. URL http://www.praat.org.

[7] SCHIEL, F.: *Automatic Phonetic Transcription of Non-Prompted Speech*. In *Proc. 14th International Congress of Phonetic Sciences (ICPhS)*, pp. 607–610. San Francisco, 1999.

[8] TEMPLETON, E., L. CHANG, E. REYNOLDS, M. LEBEAUMONT, and T. WHEATLEY: *Fast response times signal social connection in conversation. Proceedings of the National Academy of Sciences*, 119, p. e2116915119, 2022. doi:10.1073/pnas.2116915119.

[9] MAHR, T. J., J. U. SORIANO, P. J. RATHOUZ, and K. C. HUSTAD: *Speech development between 30 and 119 months in typical children II: Articulation rate growth curves. Journal of Speech, Language, and Hearing Research*, 64(11), pp. 4057–4070, 2021. doi:10.1044/2021_JSLHR-21-00206.

[10] CUCCHIARINI, C., H. STRIK, and L. BOVES: *Quantitative assessment of second language learners' fluency by means of automatic speech recognition technology. The Journal of the Acoustical Society of America*, 107, pp. 989–99, 2000. doi:10.1121/1.428279.

[11] KANY, V. and J. TROUVAIN: *Computergestützte Bestimmung des Sprechflusses bei Vorschulkindern*. In T. BAUMANN (ed.), *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2024*, pp. 62–69. TUDpress, Dresden, 2024. doi:10.35096/othr/pub-7081. URL https://www.essv.de/pdf/2024_62_69.pdf.

[12] KALLIO, H.: *The prosody underlying spoken language proficiency. Cross-lingual investigation of non-native fluency and syllable prominence*. Ph.D. thesis, University of Helsinki, 2022. doi:10.13140/RG.2.2.29682.99524.