Studies in Second Language Learning and Teaching
Department of English Studies, Faculty of Pedagogy and Fine Arts, Adam Mickiewicz University, Kalisz
SSLLT 4 (3). 2014. 665-692
doi: 10.14746/ssllt.2014.4.4.5
http://www.ssllt.amu.edu.pl

Assessing instructional effects of proficiency-level EFL pronunciation teaching under a connected speech-based approach[1]

Sasha S. Euler
University of Trier, Germany
eulers@hotmail.co.uk

Abstract

This article discusses the assessment of pronunciation instruction under a new approach to pronunciation teaching centered on the role of connected speech in the prosodic system of English. It also offers a detailed discussion of various empirical problems in teaching-oriented L2 pronunciation research and suggests ways of addressing them in intervention studies. A new explanatory sequential mixed-methods design was developed for this study, which was used to assess 10 advanced EFL learners in Germany before and after 13 weeks of instruction. The results revealed co-occurring developments in learners' use of prosody and connected speech in line with the rationale of the approach. The findings lead to various implications for language teaching and assessment. For future research, ways are suggested to increase the validity and predictiveness of L2 pronunciation research from both empirical and pedagogical perspectives.

Keywords: English language teaching, EFL, pronunciation teaching, prosody

[1] Special thanks go to J. D. Brown for his invaluable help with the mixed-methods research design and for his encouragement to go new ways in both pronunciation teaching and research. Any errors that remain are the sole responsibility of the author.

1. Introduction

Pronunciation can still be considered an “orphan” (Gilbert, 2010) in English language teaching (ELT).[2] The problem is that owing to the long neglect of this area of ELT, pronunciation pedagogy lags significantly behind the advances made in other domains of ELT. Addressing this issue, Gilbert (2010, p. 3) notes that “there need to be major changes in teacher training, materials available, appropriate supporting research, and changes in curricula,” and in Brown and Kondo-Brown (2006a) calls are repeatedly made for a *systematic approach* and the development of *practical materials* (e.g., pp. 25, 94). In order to accomplish this, a first step is to develop a view of language, that is, a view of the underlying logic of how language is structured and how its constituent parts interlock to form a meaningful whole (a good example being Lewis’ [1993, 1997] lexical approach). From there, priorities can be set and curricula and materials developed so that a workable pedagogical system may be established. This has been done in the creation of a connected speech-based approach to pronunciation teaching (ConSpA; see Euler, 2014a, 2014b, 2014c), which is the basis for the instruction conducted for the present study.

With the communicative turn in ELT, pronunciation (then understood as the drilling of isolated sounds) was initially deemed to be of little relevance, before pronunciation teaching gradually shifted to a focus on prosody owing to its inherent meaning and context-generating quality (Chun, 2002). The ConSpA (outlined in Appendix A) is a recent development of this tendency in that it centers on English rhythm and the reduction, deletion and linking processes resulting from the need to maintain stress timing. The way rhythm and connected speech create prosodic words and chunk the stream of speech into meaningful units significantly impacts students’ intelligibility, comprehensibility and accentedness if not employed adequately, and causes enormous difficulty in listening comprehension because of the mismatch between students’ representation of the language and authentic L1 speech (see Euler, 2014c for a discussion).

While intuitively plausible, empirical evaluations are essential in order to justify the alleged prioritizing of prosody-oriented teaching over segment-oriented teaching (or no pronunciation focus at all). However, such studies are very difficult to conduct because entire teaching programs need to be evaluated, which poses considerable empirical as well as logistical challenges. As a consequence, only few studies can be identified that go beyond impressionistic

[2] This situation has improved in SLA research. However, as will be argued in this paper, SLA scholars tend to follow very different paradigms so that the benefit for ELT is not clear, where pronunciation is still a much neglected area (Grant, 2014).

teacher judgments (e.g., Couper, 2003; Ricard, 1986) or artificial laboratory experimentation (e.g., Watkins, Rauber, & Baptista, 2009) and provide *both* empirically sophisticated and practically applicable results. Three of these studies will be briefly discussed.

In an earlier study, Anderson-Hsieh, Johnson, and Koehler (1992) investigated SPEAK Test raters' pronunciation judgments of readings of a text passage in relation to errors in segmentals, prosody, and syllable structure. The authors obtained general accent ratings from three experienced raters (EFL teachers who had worked as SPEAK Test raters) and statistically correlated them with a general value based on counts of individual errors marked on transcriptions for sounds, syllable structure and prosody. While sounds and syllable structure correlated significantly with the accent ratings (-.67 and -.76, respectively), prosody showed the strongest relationship (-.90) (all at $p < .0001$). However, the authors conclude that in addition to such statistical correlation data, "research is needed that further investigates the specific ways in which prosody affects pronunciation judgments" (p. 549). This need is addressed in the following studies and in the present one.

Derwing, Munro, and Wiebe (1998) conducted an instructional intervention study under the premise that more studies with longer interventions under natural conditions were needed in which the exact type of instruction is specified. To address such needs, the authors organized three different groups who received 12 weeks of instruction for segmental accuracy (1), general speaking habits and prosodic factors (2), and general English with no specific pronunciation focus (3; a placebo group). Instruction was assessed based on diagnostic sentences and extemporaneous speech data, both rated for comprehensibility and accentedness. For the sentence data a two-way ANOVA with Time (pretest-posttest) and Focus (segmental, prosodic, placebo) as factors revealed significant improvements for both pronunciation groups for comprehensibility, while accentedness improved with all three groups. The second, more naturalistic test elicited data with a picture story. Here six experienced raters rated 45-second recordings of 48 students. In this test, the accent scores proved nonsignificant, while in the comprehensibility test only the prosodic group showed significant improvements. These findings suggest that the unit of measurement (read sentences vs. extemporaneous samples) as well as the focus of instruction have an impact on pronunciation ratings. The authors argue convincingly that the high ratings for the segmental group in the sentence-reading test was a result of their undivided attention being focused on perfect production, while in the free speech test students' attention was divided and their segmental knowledge was, apparently, not transferrable, while prosodic skills were (p. 406). The transferability of prosodic skills to free speech clearly supports prosody-oriented approaches to pronunciation teaching.

Still, as ANOVA analyses are very general and can blur individual case facts, in Derwing and Rossiter (2003) the authors took the same data and conducted a more qualitative analysis to determine which forms to best focus on in teaching practice (p. 3). In this study, six professional judges rerated the recordings by marking transcriptions and commenting on prosody. Comprehensibility was classified as either problematic to comprehend, irritating, or salient (clearly noticeable) but not problematic. The authors counted all errors as either segmental/syllable structure, morphological, syntactic, semantic, filled pauses, repetition or prosodic. The authors found that "over conditions, times, and judges, errors problematic for comprehensibility [24.1% of identified errors] tended to be mostly phonological, bothersome errors [32%] were mostly due to filled pauses, and salient errors [43.7%] were predominantly morphological" (p. 10). A repeated measures ANOVA for prosody showed a significant 16% discrepancy between the global and segmental group in the posttest, confirming that prosodic instruction promotes automaticity (p. 13). The authors suggest that future research include "the description of developmental patterns in pronunciation, the effectiveness of specific activities in pronunciation instruction, and the ongoing investigation of factors that affect comprehensibility" (p. 14). The present study follows these suggestions and looks specifically at prosody and resulting connected speech phenomena.

2. Agenda of the present study

In order to assess instructional effects of the ConSpA, the following research questions were devised:

1. Can pronunciation instruction using a top-down ConSpA lead to measurably improved pronunciation proficiency?[3]
2. Which linguistic and experimental factors seem to shape or influence accent ratings?

While the first question, addressed through global accent ratings, more generally looks at the effectiveness of intervention, the second question is needed to show if the ConSpA could fulfill its purpose of making highly intricate aspects of English phonology like rhythm, elision and linking teachable (see Appendix A).

[3] Since this paper assesses highly advanced learners, intelligibility and comprehensibility are already given. The course underlying the experiment was advertised as a pronunciation and oral fluency course for advanced learners only. Since in the assessment standard language tests are used (as in the Certificate of *Proficiency* in English), pronunciation here *de facto* refers to the dimension of accentedness as at such advanced levels relative correspondence with a target norm is assessed.

For this purpose, a new mixed-methods design was employed to address Research questions 1 and 2 through its different parts and to research naturalistic language in order to maximize pedagogical applicability. Along these lines this paper also has a specific research methodological agenda. As will be shown, much of the existing L2 pronunciation research has been criticized for its lacking applicability to pedagogical contexts, so that the third questions was:

3. Which research methodology factors should be considered in teaching-oriented L2 pronunciation studies?

The rationale behind this question, starting with an extensive discussion in the following section, is to raise awareness for future research. This will be accomplished by calling on various voices from the field (see below), which may be interpreted as indicating that among scholars interested in language pedagogy, a Kuhnian state of crisis, possibly eventually leading to a paradigm shift, may be commencing (see Kuhn, 1970, 1977). Possible solutions to such empirical factors will be further discussed in the last section of this paper, which will, in addition to a discussion of the results of the experiment, also include implications for language teaching and assessment. As a consequence, this paper will be of relevance for three applied linguistic fields: instructed L2 pronunciation acquisition, ESL/EFL pronunciation teaching, and language testing. Language testing reported in this paper, as will be shown, was characterized by the fact that some language examiners were able to look beyond prosodic distortions due to nervousness and/or test/experiment conditions, while others took this as part of student performance and assessed it as such. The qualitative analysis also further validates Derwing and Rossiter's (2003) findings cited above on the effects of filled pauses, which is of further consequence for language examination purposes.

3. Research methodology factors

3.1. Empirical issues in L2 pronunciation research

Much of the published research on L2 pronunciation acquisition deals with speech perception and production. From a TEFL perspective, the problem with this research is that it is

> inaccessible to those without specialized knowledge of phonetics. Moreover, some of the research may not be perceived as practical because it has been carried out under strict laboratory conditions, so that it is not immediately clear how the findings apply to the classroom. (Derwing & Munro, 2005, p. 382)

Zampini (2008) further illustrates this issue by arguing that

> one of the primary drawbacks to laboratory-elicited speech is that it is rarely naturalistic. Indeed, some studies elicit the production of isolated syllables, words, or short phrases out of context, and the results of such studies may not be generalizable to speech in a natural setting. Some even question whether or not conversations conducted in a laboratory setting can be considered truly spontaneous or naturalistic. (p. 239)

The statement in the last sentence will be addressed in detail below and in the discussion section since facilitating the ability to use language spontaneously and in naturalistic settings should, arguably, be the aim of any L2 teaching effort.

In addition, and especially as regards instructional intervention studies, there are several design features relating to what is assessed that need to be addressed systematically. These features can be summarized under the following bullet points:

- competence/acquisition versus performance
- extemporaneous/spontaneous production versus controlled/self-monitored production
- implicit or automatized versus explicit knowledge
- large versus small scope and duration
- communicative/interactive instruction versus teacher-centered or computer-based instruction
- applied linguistics versus linguistics applied

While in experimental phonetics/phonology or in more general linguistics-oriented SLA studies naturalistic language may not always be of relevance, in L2 phonological studies designed to inform language teaching (such as instructional intervention projects), the analysis of natural (or authentic/spontaneous/extemporaneous/conversational/free) speech seems necessary since, as has been said, producing such language is the explicit aim of any L2 instruction. This also has a strong psycholinguistic rationale. In the context of speech perception, Strange and Shafer (2008, p. 166) argue that "real spoken word recognition requires rapid identification or categorization of phonetic segments by reference to internalized representations of those categories." In experiments that are highly controlled, students are very likely to show what they are intellectually capable of, rather than what they have internalized, that is, what they would be capable of when the cognitive load is on communication (as has been indicated in the above Derwing et al. studies). In the context of intonation, Chun (2002, pp. 89-90, 94) addresses this problem from a useful perspective. She argues that a "distinction has to be made between acquisition phenomena and performance phenomena" since, especially in the context of research supporting ELT, "simply being able to demonstrate aspects of intonation physically and perceptually does not necessarily render the process useful from a teaching point of view." Internalized knowledge is not demonstrated adequately through "mere" performance, so that empirical research should be adapted accordingly. Likewise, Ellis (2013) argues from an instructed SLA perspective that while free constructed responses (communicative tasks) often empirically show little effect, they "arguably . . . constitute the best measure of learners' L2 proficiency" (p. 41).

Such arguments clearly relate to the type of knowledge demonstrated in elicited data. In looking back at meta-studies on the effectiveness of explicit instruction, Spada (2011) remarks that the alleged advantage of explicit instruction may be

> due to the fact that the majority of tests to measure learners' progress in instructed SLA research are tests of explicit knowledge. In fact, 90% of the outcome measures in the primary studies included in the Norris & Ortega [2000] meta-analysis used highly constrained, discrete-focus linguistic tasks, whereas only 10% required extended communicative use of the L2. (p. 228)

While building explicit or declarative knowledge can speed up the acquisition process, and while automatized procedural knowledge *can* become "functionally equivalent" to implicit knowledge (DeKeyser, 2003, pp. 329-330), the inert knowledge problem (the nontransferability of grammar knowledge to free speech) must not be ignored (Larsen-Freeman, 2003). This is, in fact, a general problem in instructed SLA research. In the context of the implicit-explicit learning issue, for example, DeKeyser (2003, p. 336) argues that L2 research tends to be quite limited because studies are *too small in scope and duration*, they are *conducted in laboratory settings rather than classrooms* and *criterion measures tend to be very constrained* (grammatical judgment or fill-in-the-blank tests vs. freely constructed discourse). These factors hold equally true for L2 pronunciation research. There are many studies in which students receive some kind of computer training in the perception or production of minimal pairs, or a few brief frontal (teacher-centered/lecture-type) instruction sessions on articulatory settings and are then asked to read out or repeat wordlists in a laboratory, which is supposed to demonstrate their acquisition of English pronunciation (see e.g., some of the studies in Watkins et al., 2009). However, as has been argued by Chun (2002), such studies most likely demonstrate perceptual or productive *performance*, rather than actual *acquisition*. Owing to all such aspects, this kind of methodology renders such studies nearly completely irrelevant from a teaching perspective. In ELT it is well known and needs no further elaboration that language acquisition is best facilitated through meaningful interaction within a larger language program, and that language cannot be "taught" as in transferred from teacher to student (e.g., H. D. Brown, 2007; Nunan, 2004). Therefore, the type of instruction in an experiment should also reflect the nature of classroom second language learning.

Along similar lines, Piske, Flege, MacKay, and Meador (2011) were able to confirm the hypothesis that errors observed in studies on sounds may well be only artifacts of the elicitation technique employed. The authors argue that while "most researchers would acknowledge that conversational speech should represent the most important criterion for success in acquiring L2 vowels . . . surprisingly few studies have been undertaken," which is most likely due to "the inherent difficulty in analyzing conversational speech under controlled conditions" (p. 2). This question concerns the balance between the ecological validity of a study (how similar the experiment is to natural communication or settings) and experimental control to achieve token richness and comparability (Post & Nolan, 2011; see also DeKeyser, 2003). The Derwing et al. studies reviewed above and the new mixed-methods design presented here explicitly try to strike such a balance. However, they still "simply" follow customary methods in L2 studies. In the discussion section, ways of adapting methods from other traditions for L2 research will be suggested.

As has been indicated, the fact that such discrepancies are discussed in review articles and handbook chapters such as the ones cited may be of significance from the perspectives of applied linguistic theory (e.g., Davies, 2007; Widdowson, 2000) and, especially, Kuhnian philosophy of science (Kuhn, 1970, 1977), indicating possible commencing paradigm shifts. Many years ago Henry Widdowson attempted to set a course for applied linguistic research with his contrast between a theoretical notion of "linguistics applied" and actual "applied linguistics":

> With linguistics applied the theory of language and the models of description deriving from it must be those of linguistics. As an activity, therefore, it is essentially conformist. . . . For applied linguistics, the central question is: How can relevant models of language description be devised, and what are the factors which will determine their effectiveness. (Widdowson, 1984, p. 22)

Much of the work criticized above is essentially conformist in that the focus is on the linguistic tools available and how they can be used, rather than on real-world needs which can be identified and addressed (cf. Dörnyei, 2007, p. 17). In other words, while applied linguistics (such as ELT/TESOL) would start with classroom needs, linguistics applied (or simply general linguistics) aims at identifying established research paradigms and starts with scientific needs as reflected in these paradigms. The consequence of this is that in purely phonetic/phonological or general linguistic SLA studies the features criticized above may simply be considered the way "normal science" (to use Kuhn's terminology again) is practiced under established paradigms. In teaching-oriented studies, however, it is necessary to start with real-world needs such as the nature of classroom language acquisition and teaching. This will, then, often create a different view of which issues still need to be solved, so that methodology has to be adapted to these needs (Davies, 2007, p. 29).

While more general linguistics/phonetics-oriented studies would often go for the second options listed in the above bullet points, the present study (even though it was quite restricted by institutional conditions) attempts to realize the first options as they are considered most conducive to pedagogical applicability.

3.2. The mixed-methods research (MMR) paradigm

Since the mixed-methods paradigm is still rarely found in L2 pronunciation research, it seems important to highlight some theoretical underpinnings of this paradigm and the particular variant of it employed here, before design and results of the experiment are elaborated on. One of the distinctive purposes of mixed-methods research is that it can aid in *achieving a fuller understanding of a target phenomenon* and in *verifying one set of findings against another*. This is achieved by "trying to make 1 + 1 = 3 by carefully combining qualitative and quantitative data and analyses" (Brown, in press b, Chapter 6). While there are several different types of MMR designs, the present study uses an "explanatory sequential" design (Cresswell & Plano Clark, 2011). In this design, the data collection procedures involve first collecting quantitative data, then analyzing the data, and using the results to inform the follow-up qualitative data collection (Cresswell & Plano Clark, 2011, p. 185). This procedure is useful because researchers can interpret how the qualitative results help to *explain* the initial quantitative results. In practice, this is accomplished by selecting interesting, illustrative or odd cases from the whole population studied in the quantitative part for the qualitative follow-up analysis (Cresswell & Plano Clark, 2011, p. 181), which is thus not to be confused with "cherry picking" (compare Dörnyei, 2007). Especially as regards classroom SLA and pretest/posttest intervention studies, this is particularly useful because various secondary and developmental factors may play a role at a given moment. Possible issues of external validity aside, one consequence of this design can be that not all students may actually demonstrate their regular performance during data elicitation (e.g., Nunan, 2004, p. 30). The selection of cases for the qualitative follow-up analysis in this kind of MMR is a useful way of counteracting this problem.

4. Evaluation of the ConSpA

4.1. Design

The explanatory sequential design is realized in the present study as follows:

1. Phase I, quantitative data collection: Five professional Cambridge English Language Assessment speaking examiners from Britain and the USA were employed to assign global accent ratings, classified in a numerical system on a continuous nominal judgment scale. However, in order to avoid ceiling effects, as only advanced/proficiency learners were assessed, the Cambridge benchmark system realizing the Common European Framework (CEFR) for the levels B2, C1 and C2 (high intermediate, advanced, proficiency) was employed, resulting in a 9-point scale. Appendix B presents a variation of the Cambridge realization of the CEFR levels as it was given to the judges (further discussed in 4.3.1).
2. Phase II, qualitative data collection: A number of coding schemes (inspired by Brown & Kondo-Brown, 2006b) and semantic scales were devised in order to obtain a view of prosodic, segmental and connected speech deviations or realizations and of formal mistakes (compare Derwing & Rossiter, 2003). The data (recordings of oral production) were analyzed combining auditory analyses by the author and another trained linguist with ELT experience, and the use of phonetic speech analysis software.

4.2. Procedure

4.2.1. Instruction and participants

The instruction was conducted at a private language institute and at a university in Germany (13 90-min sessions over 15 weeks) as a pronunciation and oral fluency course under a ConSpA framework following a North American English model. The underlying idea of the ConSpA is that connected speech becomes processable and therefore teachable by establishing it as a logical consequence of the workings of the prosodic system, resulting in a teaching sequence as illustrated in Appendix A. Methodologically, this course followed the Celce-Murcia model of communicative pronunciation teaching (Celce-Murcia, Brinton, & Goodwin, 2011, pp. 44-48), but also included task-based units (see Euler, 2014c). 10 students in two groups (7 German, 2 Italian, 1 Polish, all between their 20s and early 40s) attended regularly and agreed that their recordings be used for research purposes. Students reported that they were not receiving any other instruction and that they did not have the time for any additional practice at home. In addition, there was a control group of 4 native speakers from the USA and Canada whose recordings were assessed alongside students' recordings. As raters were not aware of the fact that some recordings came from native speakers, these ratings were used to confirm that the highest levels were, indeed, assigned appropriately.

4.2.2. Collection of listening stimuli

In pre- and posttests, conducted on the first and last day of the course, diagnostic passages and a picture elicitation task (adapted from a past C1 Cambridge ESOL examination) were employed. However, the oral reading data could not be used since reading with authentic prosody and connected speech is extremely challenging even for native speakers, resulting in unnatural and distorted speech (see also the discussion section). The rationale behind the free speech task was to increase the cognitive load as appropriate for advanced learners so that no or only very little monitoring should be possible. The picture elicitation task (in which students were asked to discuss how realistic they consider certain dreams, which were illustrated, to become reality) was timed and took approximately 90 seconds, though the exact number of words differed slightly due to speech rate and hesitation.

In both tests students spoke into a Logitech H555 microphone. The speech samples were digitally recorded with 16-bit. For the accent ratings all 69 resulting recordings (oral reading and free speech tasks including the control group) were randomized and raters were instructed to assign a value between 1 and 9 as elaborated on an accompanying sheet with CEFR benchmark descriptors for the numerical values (see Appendix B). The raters were also given three warm-up items, one from a native speaker and two from students, to familiarize themselves with the type of advanced English they were going to assess.

4.3. Quantitative data

4.3.1. Analysis

Instead of using the typical continuous rating scales going from, for example, *very strong accent* to *no accent* (see e.g., Derwing & Munro, 2005; Flege, Munro, & MacKay, 1995), for this study a CEFR-based scale was developed since this should make it easier for trained teachers or language examiners to produce objective and reliable ratings. It also provides a detailed system for assessing specifically advanced English. Creating such a detailed assessment system was a high priority because, as previous research has shown (Bongaerts, vam Summeren, Planken, & Schills, 1997), assessing pronunciation in advanced and proficiency

English has been found to be a very difficult task (as the raters in the present study also reported). Consequently, a type of indicator triangulation was used here that combines the general continuation of the scale (1 through to 9) with the familiar CEFR levels (B2, C1, C2) and specific level descriptors for each level based on the Cambridge English Language Assessment benchmark descriptors (see Appendix B). However, since Cambridge does not have separate descriptors for Levels C1 and C2, their benchmark descriptors were modified to account for differences on those levels. The raters were instructed to assess the pronunciation dimension of the recordings just as they would in any regular Cambridge exam (including the recordings of the native speaker control group, of which they were unaware), but were instructed to consider the sheet given in Appendix B as it modified the level descriptors somewhat. The native speaker control group was included to demonstrate that full proficiency can, indeed, be identified. Likewise, in the free speech test all native speakers received Level 9 ratings.

Inter-rater reliability was calculated via SPSS using a two-way mixed intraclass correlation coefficient (ICC) in order to calculate reliability across multiple raters and multiple informants. Inter-rater reliability for the free speech task was *r* = .72 (calculated for ultimate agreement), which reflects moderate reliability as often found in such studies. The fact that even experienced expert raters could not achieve a higher level, again, is explained through the fact that only advanced English was assessed in a highly detailed system (while typically rating scales cover the whole spectrum).

4.3.2. Results

Table 1 contains mean values showing developments from pretest to posttest. In order to interpret these ratings, a number of secondary factors relating to language assessment need to be considered.

Table 1 Accent ratings in the pretest and posttest

Student	Free speech task *M*s	
	Pretest	Posttest
1	5.6	6.3
2	3.0	3.8
3	4.8	5.2
4	5.2	6.0
5	6.8	7.6
6	7.2	7.3
7	5.0	6.4
8	6.2	7.2
9	2.4	1.8
10	3.0	3.6

A very strong influencing factor is the psychological stress (as all students openly reported and as was quite obvious from their behavior) of being in a kind of exam situation and being recorded for error analysis (even though verbal attempts were made to ease the stress). It should be said in this context that this was a voluntary class and not subject to high-stakes examination, so the reported test was entirely for the purpose of the experiment (and students' desire to know if they improved).

In the pre- and posttests some students were doing notably worse than in classroom interaction, where Student 6, for example, regularly showed native-like production. The linguistic factors resulting from this kind of stress (unnatural duration, pitch and loudness, hesitation phenomena distorting rhythm and intonation contours) will almost necessarily have an impact on accent ratings. In connection to this, it is rather obvious that in such experimental situations some connected speech phenomena are unlikely to be employed as they would in conversational speech (Warner, 2011), so ratings have to be treated with care. Further, low ratings could be a consequence either of lacking acquisition (Student 9, whose starting level was so low that the instruction was beyond his zone of proximal development and thus not processable [e.g., Pritchard & Woollard, 2010]), or of random fluctuations in performance (Student 6, whose performance was generally quite native-like but who would occasionally lag behind her regular level, thus showing that her interlanguage system had not fully stabilized). While this study used standard procedures in order to explore and confirm expected limitations in L2 pronunciation research (the reading passages will be discussed below), in the last part of this paper some possibilities (quite unusual for L2 studies) will be elaborated on that may serve to improve this situation. An additional very strong factor is rater subjectivity (cf. Munro, 2008). This factor is two-fold: first, it concerns what "bothers" or "impresses" individual raters more or less, and, second, what raters happen to pay attention to at a particular moment in time. Such aspects explain how even experienced expert raters were only able to achieve moderate reliability and has clear implications for language teaching and testing.

4.4. Qualitative data

4.4.1. Analysis

As has been elaborated above, the accent ratings serve as the basis for both case selection and further cross-verification in the explanatory-sequential MMR design. While accent ratings provide a global score of native speakers' reactions, it will need the detailed qualitative analysis to determine which exact features might have shaped those ratings.

In this part, 4 students were selected for detailed analysis because they posed particularly interesting cases in different ways. Specifically, from the teacher's perspective, Student 4 did not seem to have developed significantly throughout the course, which is surprising. Student 5 improved from close to Level 7 to close to Level 8 (receiving several Level 8 ratings in the posttest), resulting in clearly native-like performance. Student 7 did develop somewhat throughout the course, but the increase by 1.4 bands is surprising since many deviations still seemed to be present. To explain such cases, a thick qualitative description interpreted on a case to case basis will allow making assumptions as to whether certain (combinations of) features seem to have a positive or negative bearing on accent perceptions. Still, such assumptions, however informed, are always just that, and would need to be replicated with different students and different raters in different scenarios. For such a qualitative data analysis to be possible in a sufficiently controlled and replicable manner, detailed coding schemes need to be developed (J. D. Brown, personal communication, November 2012), here showing precisely how many realizations out of how many possible realizations could be found in the pretest and posttest recordings. The prosody scale classified certain phonological features on a 3-point semantic scale. This system, in the form of category goodness, is widely used for sounds in L2 phonological studies and was also employed here for prosody.

As regards data analysis procedures, printed transcriptions (in normal spelling) of all the extemporaneous speech samples (pretest and posttest) were produced, with phonetic transcriptions by the author of possible vowel reduction, assimilation and deletion, as well as linking processes and allophonic variation (cf. Brown & Kondo-Brown, 2006b). These transcriptions illustrate the amount of *possible* realizations in each recording, leading to the number after the slash in the following tables, and were thus done only with the text, that is, without listening to the recordings. Further markings were used for pitch raises or drops and for very likely tone unit boundaries. The latter was important since in this way certain words or phrases could be marked as cases where connected speech processes were unlikely owing to prosodic factors (first word of a tone unit without notable anacrusis, relative sentence stress). With assimilation, virtually no realizations could be observed because being asked to speak into a microphone in order to be recorded for language assessment is not the most natural situation for using a lot of such connected speech reductions (Warner, 2011), although other processes such as vowel reduction to schwa and non-assimilation types of linking are largely a fixed part of (L1) English.

4.4.2. Results

Table 2 shows results productive for connected speech (several additional types of assimilation, elision and linking were counted but did not yield a sufficient number of tokens). Examples include *h* or *th* deletion in pronouns or deletion of *d* in *and*, vowel reduction to /ə/ (schwa) in unstressed function words (*of, from, them, that, can, you, some,* etc.), and linking with [j] and [w] (*blu^{w}eyes, see^{j}it*). It should be said here, however, that the exact *number* of connected speech reductions used is always, to some extent, an idiosyncratic decision.

Table 2 Connected speech realizations (the number of realizations/the number of possible realizations)

Student	Deletion in function words		Vowel reduction		Glide insertion (vowel + vowel linking)	
	Pretest	Posttest	Pretest	Posttest	Pretest	Posttest
1	3/8	9/14*	5/12	10/15*	1/5	8/11*
4	5/10	2/6	7/23	5/13	4/10	5/6*
5	5/12	7/9*	4/15	14/24*	4/6	7/9
7	3/11	7/10*	2/19	10/21*	5/7	5/8

Note. *Improvement over the pretest.

For prosody a 3-point system emulating category goodness in sound studies was used with a high, intermediate and low value (for advanced learners) for problematic aspects that are reliably measurable combining auditory judgments and instrumental analysis. Rhythm was rated on a 3-category semantic scale as *natural* (3), *sometimes distorted* (2) or *often distorted* (1) (in L2 speech also due to lacking fluency or semantic gaps). Such a rather subjective assessment is needed because "speech rhythm is very much a perceptive category. We perceive speech as rhythmic even if the physical placement of articulatory events does not occur at precisely regular time intervals" (Szczepek-Reed, 2011, pp. 140-141). This is further supported by the fact that rhythm is often broken due to hesitations or the dynamics of conversational interaction, so that regular speech rhythm typically only occurs over stretches of two or three intonation phrases at most (p. 146).

Vowel duration and pitch in intonation units were assessed as the main markers of sentence stress. Vowel duration can not only be easily measured instrumentally but is essential in marking (primary) sentence stress and can be missing as a stress marker even if pitch and loudness are employed naturally (Chela-Flores, 1997). Pitch is rated in its main domain, the intonation unit. An intonation unit has an overall intonation contour, at least one pitch accent (elements receiving sentence stress) and typically one primary/nuclear pitch accent (with a higher pitch level). Such features can, again, be measured quite reliably combining auditory and instrumental analyses (as portrayed in Szczepek-Reed, 2011). While discussed on a case-to-case basis in Section 4.1, Level 1 ratings were used with significant distortion across the band. Level 2 signifies that many correct tokens could be observed but that distortions are still present, while Level 3 shows native-like correctness/consistency. In the special case of Student 7, no clear value could be determined, as discussed in detail below.

Next, filled pauses and repairs (sentences that are aborted in the middle and restarted) are measured as occurring in *high amounts* (1), *occasionally* (2), or in *low amounts* (3). As it has been shown (e.g., Derwing & Rossiter, 2003) that filled pauses can cause annoyance in listeners (as the raters in the quantitative part agreed when asked afterwards), this aspect was also measured impressionistically to approximate real-world listeners' perceptions.

Table 3 Category goodness in prosody

Student	Rhythmic timing		Vowel duration in sentence stress		Pitch in tone units		Filled pauses/ repairs	
	Pretest	Posttest	Pretest	Posttest	Pretest	Posttest	Pretest	Posttest
1	3	3	2	3*	2	2	1	2
4	2	3*	3	3	2	2	1	3*
5	3	3	3	3	2	3*	2	2
7	2	2-3*	2	2-3*	2	2-3	1	1

Note. *Improvement over the pretest.

Finally, sounds (certain vowels and consonants and some allophonic variation) and formal errors were also counted. Since sounds were not an essential part of the instruction (only small portions of the whole program were devoted to segmentals out of prosodic contexts, following the rationale of the studies cited in the introduction and current trends in ELT) not many distinctive improvements occurred. As regards formal errors, only grammar was consistently productive, showing between two and four mistakes in each recording.

5. Discussion

5.1. Discussion of the results

Instructed language acquisition goes along at least three phases of measureable performance manifestations. A first phase centers on noticing and beginning integration. A second phase comprises restructuring and fine-tuning of the interlanguage (IL) system, while in the final stage the IL system is stabilized through deeper processing and automatization so that new features can be used both accurately and fluently (de Graaf & Housen, 2009, p. 731). Immediately post-instruction, students are likely to be in a stage in which their IL is destabilized and integration and refinement were initiated, but it is unlikely that full integration has already been achieved. This poses an empirical problem in that results of immediate posttests are likely to be rather moderate. Still, even moderate improvements, especially when demonstrated through in-depth analyses, such as those used in the present study, can point toward the effectiveness of instruction by showing that acquisition was initiated and is in progress (Long & Robinson, 1998, pp. 40-41). Keeping such general reservations in mind, it could be shown that both the quantitative and qualitative analyses revealed clear developments. In addition, while delayed posttests were not institutionally possible, email contact with students hints at further developments based on self-assessment and current instructors' comments on their pronunciation.

As regards explaining the accent ratings through the qualitative findings, obviously only educated assumptions can be made and further cross-validation would be needed. Student 1 made minor improvements across the board, but especially his pitch was still clearly non target-like, though, quite notably, he was now able to consistently use duration (vowel length) to realize primary sentence stress. Student 4's ratings were surprising (thus his selection) because he appeared quite resistant to instruction during the intervention. This student improved significantly in his use of rhythm and filled pauses, the latter being naturally related to the former. This further validates Derwing, Munro, and Wiebe's (1998) method of approaching prosodic instruction as "global" focus on prosody and speech habits, and it underlines the recent pedagogical prioritization of rhythm in pronunciation teaching (Brown, 2013; Brown & Kondo-Brown, 2006a; Cauldwell, 2013; Euler, 2014b, 2014c). Even though Student 5 was already very advanced from the beginning, she achieved a largely native-like level, which is a significant achievement in its own right. She significantly improved in all categories, very notably so in vowel reduction and pitch, so probably only minor accent traces and very few grammatical errors and hesitation phenomena gave her away as an L2 speaker. Student 7 was selected because she showed a very high amount of hesitation in both tests and seemed to have clearly not mastered many phonological features. Despite that, her accent ratings showed higher improvements than those of any other student, although they also showed more inter-rater variation than anyone else's. The in-depth analysis revealed that some prosodic features improved but were still not entirely target-like, and that clear developments in connected speech realizations were made. While still very much in progress, the improvements in prosody and connected speech seemed to have made a better and significantly more native-like impression on several raters, while others were probably disturbed by the still high amount of hesitation (even though more prosodically aligned) and/or individual negative tokens. The qualitative analysis of the posttest data did not reveal a distinctive score (either Level 2 or 3) as it was very difficult to distinguish between hesitation-induced prosodic deviations and lacking competence. In the same way, inter-rater reliability for Student 7 in the accent ratings was very low, with individual ratings in the posttest varying between values 4 and 8 (they are further interpreted in Section 4.3).

5.2. Discussion of empirical issues

It was part of the agenda of this paper to illustrate and discuss empirical issues in teaching-oriented L2 pronunciation research. One of the primary concerns here was if read-out language can validly be used for such research. While the studies in the introduction already established that the unit of measurement has a strong impact on language data, this study confirmed possible reservations against read-out language owing to the distortedness of the reading passage data. This was intentionally taken to extremes with a connected speech dialogue (taken from Brown & Kondo-Brown, 2006b), where (experienced expert) raters differed by up to 7 points on the same recording. Student 6, for example, received the following five ratings in the pre- and posttest, respectively: 4, 4, 6, 3, 4 versus 8, 1, 3, 2, 8, which was the most extreme case as some raters considered it completely substandard while others considered it to be native-like. This suggests that prosody and connected speech cannot be usefully studied this way (compare also the section on assessment below). While in general SLA research it has been argued that ecological validity (here, experiments being conducted just as regular classroom teaching) should not be the "sacred cow" (compare DeKeyser, 2003, p. 339), for teaching-oriented pronunciation research it should, indeed, be of primary concern (possible techniques are discussed presently).

In a similar manner, more generally the question was posed "whether or not conversations conducted in a laboratory setting can be considered truly spontaneous or naturalistic" (Zampini, 2008, p. 239) at all. Indeed, it could be observed that some students' test performance tended to significantly lag behind classroom performance, again especially so in the domains of prosody and connected speech. A possible solution that would be worth exploring (and that could also be used to empirically show the difference between prosodic test and classroom performance) may be to obtain data from classroom interaction without students' knowledge (with their general consent, of course) while engaged in task work. The resulting interactive data could then, for example, be analyzed with the system developed for the study of pronunciation in interactional linguistics and conversation analysis (e.g., Couper-Kuhlen, 2007; Szczepek-Reed, 2011).[4]

The other criteria introduced above (testing acquisition/competence and implicit/automatized knowledge, running large-scope long-term interventions, and using communicative and interactive instruction) were also specifically addressed in this study. The first two issues were tackled through the use of C1 oral exam tasks creating a level-appropriate cognitive load that would only allow for minimal monitoring and would test transferable and internalized knowledge (see the references above). Warner (2011), speaking from a laboratory phonology perspective, however, lists this as only one out of at least seven workable methods for eliciting natural language rich in connected speech, in the discussion of which he also explicitly cautions against the effects of physically being in a laboratory setting and having to wear or speak into a microphone since this can make language less natural (and therefore also less rich in reduction processes), as observed in this study. The author illustrates several factors of immediate relevance here that can help to make speech more natural and rich in reduction: *engaging in a dialogue* with a moderator or an interlocutor, the *presence of familiar interlocutors* (family, friends), or *being recorded without immediate awareness* of it. In L2 teaching contexts, these criteria could be realized with classmates as conversation partners or the teacher as a moderator, though the unaware classroom interaction recordings, as already suggested, may represent the most natural condition by far. With this technique some control would naturally be achieved by having students engage in the same kind of task work. If these tasks have a specific pronunciation focus (e.g., identifying instances of sentence stress patterns in a text in order to work with them further (see Nunan, 2004), even more control would be exerted quite organically (though this should be combined with entirely meaning-focused tasks as monitoring will always be present even in meaningful and interactive structure-oriented tasks). In either case, it would be very much worth exploring such techniques in future L2 pronunciation research.[5]

6. Implications for language teaching and assessment

As regards teaching, it has been said in the introduction that the ConSpA is a development of the prosody-centered state of the art in pronunciation teaching

[4] Note that Cambridge English Language Assessment also uses student interaction as one of their primary elicitation techniques in language tests.

[5] As a further aspect, large-scale intervention studies and pretest-posttest designs have been criticized on more general grounds by SLA scholars (as mentioned above). Since this paper has a more practical and specifically pronunciation-oriented focus, this discussion will not be reviewed here.

(e.g., Chun, 2002; Gilbert, 2010, 2012) with the specific aim to help learners achieve authentic production and comprehension. However, a major problem with aspects like rhythm and connected speech is their intangible and elusive nature, so these aspects of the English phonological system are extremely difficult to make processable for students and to teach in a communicative, interactive and motivating manner (cf. Euler, 2014a, 2014b). The present study has shown that developments in all domains were observed while none of the criteria of good pedagogy needed to be neglected. In Euler (2014c) it has, in addition, been shown how the way rhythm and connected speech structure the stream of speech into meaningful chunks provides a very fruitful gateway into task-based pronunciation teaching. As a consequence, the study of task-based pronunciation teaching under the premises discussed in this paper seems a fruitful direction for future research.

On a more general level, the empirical problems with the reading passage, especially in the context of the results from the Derwing et al. studies cited above, also have distinctive pedagogical implications. Derwing and her colleagues have shown that while students were able to show clear developments in the reading passage in the sound domain, only prosodic skills were transferrable to free speech. Pedagogically, reading out as a pronunciation teaching technique, indeed, only leads to reading without internalization (Celce-Murcia et al., 2011, p. 11). In fact, it is quite possible to get even beginning learners to read out a sentence with nearly perfect pronunciation, but this has no impact whatsoever on their spontaneous language use, that is, on their IL system. Such findings suggest that reading out is a cognitive skill very different from pronunciation. The flip-side of this issue, as this study has shown, is that reading out for authentic prosody and connected speech is extremely difficult even for native speakers (who often only achieved Level 7 or 8 ratings too in their readings of the diagnostic passages). This is not only highly significant for empirical research but shows again that reading passages are neither a useful tool for pronunciation teaching nor for assessment.

As regards assessment specifically, recently one of my colleagues reported what was to her a rather illuminating event when I gave her some native speaker readings of a diagnostic passage that she had used to assess her students' pronunciation performance. She was dissatisfied with her students' acquisition of prosody and thought her teaching to have been ineffective. She was rather surprised when she found that native speakers also made several "mistakes," which finally convinced her that reading out is not a valid tool for pronunciation assessment. Another important implication for language testing is the influence of hesitation phenomena. Phonologically, hesitation can break rhythm, which will automatically also impact the marking of primary sentence stress and the use of connected speech. In short, hesitation phenomena can completely disrupt prosody in English; *can* is used here because native speakers often align hesitation phenomena rhythmically (Szczepek-Reed, 2011, p. 151) or use appropriate repair strategies. With the latter point, however, language analysis moves from pronunciation to discourse management, and these two domains are assessed separately in official language examinations. It has been argued that while some judges in the present study were able to look beyond prosodic distortions due to nervousness (as students reported) and/or test/experiment conditions, others took this as part of student performance and assessed it as such. This was hinted at in Derwing et al.'s studies, and it showed specifically in the assessment of Student 7's results as discussed. It can be hypothesized that while some raters took hesitation as a discourse phenomenon and therefore "accepted" prosodic distortions, others took them as phonological deviations, with some raters probably doing both to varying degrees, thus arriving at more intermediate ratings. This is clearly something language examiners should be made aware of.

7. Conclusion

Derwing, Munro and Wiebe (1998, p. 408) noted in the conclusion to their study that "given the growing emphasis on pronunciation as of late, we look forward to a clearer empirical identification of useful and effective approaches." This was attempted through the ConSpA, working under the rationale that utilizing the organic interplay between certain prosodic features and connected speech processes makes these areas of phonology processable and teachable in a motivating, communicative and interactive fashion. While the quantitative analysis of this study showed that students, generally, made progress, the qualitative analysis was able to reveal that students did in fact develop in the intangible domains addressed through the ConSpA, and that positive effects on native listeners can be observed. Still, caution is needed since—also owing to obvious logistic factors—this study was somewhat exploratory in nature and aimed at analyzing only a limited number of students in detail. Further replication would be needed, which would especially benefit from long-term delayed posttests if this is logistically feasible. The present study, further, identified a number of empirical problems often inherent in applied linguistic pronunciation studies, especially as tests of the effectiveness of L2 instruction—the practical realization of "approaches"—is concerned. The purpose of this was to build awareness of these issues and to offer possibilities of addressing them systematically.

Several additional research directions are suggested by this study. On a research methodology level, it would be beneficial to further explore the use of mixed-methods research in teaching-oriented L2 pronunciation research. This may be especially useful since MMR lends itself to the study of naturally occurring language as relevant in tests of learner proficiency while still maintaining an appropriate level of empirical control and predictiveness. On a more practical level, teaching-oriented L2 pronunciation research still needs to further explore areas such as developmental patterns in instructed pronunciation acquisition, the interplay between various linguistic features in pronunciation teaching and learning, and their effect on native listeners. As regards the ConSpA system, obviously further tests in different settings would be beneficial in order to potentially further validate its effectiveness and implementability in different contexts.

References

Anderson-Hsieh, J., Johnson, R., & Koehler, K. (1992). The relationship between native speaker judgments of nonnative pronunciation and deviance in segmentals, prosody, and syllable structure. *Language Learning*, *42*, 529-555.

Bongaerts, T., van Summeren, C., Planken, B., & Schills, E. (1997). Age and ultimate attainment in the pronunciation of a foreign language. *Studies in Second Language Acquisition, 19*, 447-465.

Brown, J. D. (2013). *New ways in teaching connected speech*. Alexandria: TESOL.

Brown, J. D. (in press b). *Mixed methods research for TESOL dissertations and practice.* Edinburgh: University of Edinburgh Press.

Brown, J. D., & Kondo-Brown, K. (Eds.). (2006a). *Perspectives on teaching connected speech to second language speakers.* Honolulu, HI: University of Hawaii Press.

Brown, J. D., & Kondo-Brown, K. (2006b). Testing reduced forms. In J. D. Brown & K. Kondo-Brown (Eds.), *Perspectives on teaching connected speech to second language speakers* (pp. 247-265). Honolulu: University of Hawaii Press.

Brown, H. D. (2007). *Teaching by principles: An interactive approach to language pedagogy* (3rd ed.). White Plains, NY: Pearson Education.

Cauldwell, R. (2013). *Phonology for listening: Teaching the stream of speech*. Birmingham: SpeechinAction.

Celce-Murcia, M., Brinton, D. M., & Goodwin J. M. (2011). *Teaching pronunciation: A course book and reference guide* (2nd ed.). Cambridge: Cambridge University Press.

Chela-Flores, B. (1997). Rhythmic patterns as basic units in pronunciation teaching. *ONOMAZEIN*, *2*, 111-134.

Chun, D. M. (2002). *Discourse intonation in L2: From theory and research to practice*. Amsterdam: John Benjamins.

Couper, G. (2003). The value of an explicit pronunciation syllabus in ESOL teaching. *Prospect, 18*, 53-70.

Couper-Kuhlen, E. (2007). Situated phonologies: Patterns of phonology in discourse contexts. In M. C. Pennington (Ed.), *Phonology in context* (pp. 186-218). Basingstoke: Palgrave Macmillan.

Creswell, L. W., & Plano Clark, V. L. (2011). *Designing and conducting mixed-methods research* (2nd ed.). Thousand Oaks: Sage.

Davies, A. (2007). *An introduction to applied linguistics: From practice to theory* (2nd ed.). Edinburgh: Edinburgh University Press.

de Graaf, R., & Housen, A. (2009). Investigating the effects and effectiveness of L2 instruction. In M. H. Long & C. J. Doughty (Eds.), *The handbook of language teaching* (pp. 726-755). Malden, MA: Wiley-Blackwell.

DeKeyser, R. (2003). Implicit and explicit learning. In C. J. Doughty & M. H. Long (Eds.), *The handbook of second language acquisition*. (pp. 313-348). Malden, MA: Wiley-Blackwell.

Derwing, T. M., & Munro, M. J. (2005). Second language accent and pronunciation teaching: A research-based approach. *TESOL Quarterly, 39*, 379-397.

Derwing, T. M., Munro, M. J., & Wiebe, G. E. (1998). Evidence in favor of a broad framework for pronunciation instruction. *Language Learning, 48*, 393-410.

Derwing, T. M., & Rossiter, M. J. (2003). The effects of pronunciation Instruction on the acquisition, fluency, and complexity of L2 accented speech. *Applied Language Learning, 13*, 1-17.

Dörnyei, Z. (2007). *Research methods in applied linguistics*. Oxford: Oxford University Press.

Ellis, R. (2013). Principles of instructed second language learning. In M. Celce-Murcia, D. M. Brinton, & M. A. Snow (Eds.), *Teaching English as a second or foreign language* (pp. 31-46). Boston: Heinle Cengage Learning.

Euler (2014a). Implementing a connected speech-based approach to pronunciation teaching. In T. Pattison (Ed.), *IATEFL 2013 Liverpool conference selections* (pp. 104-106). Canterbury: IATEFL.

Euler (2014b). Approaches to pronunciation teaching: History and recent developments. In J. Szpyra-Kozłowska, E. Guz, P. Steinbrich, & R. Święciński (Eds.), *Recent developments in applied phonetics* (pp. 35-78). Lublin: Wydawnictwo KUL.

Euler (2014c). From communicative to task-based pronunciation teaching: Utilizing the power of rhythm and connected speech. *Speak Out! 51*, 5-15.

Flege, J. E., Munro, M. J., & MacKay, I. R. A. (1995). Factors affecting the strength of perceived foreign accent in a second language. *Journal of the Acoustical Society of America*, *97*, 3125-3134.

Gilbert, J. B. (2010). Pronunciation as orphan: What can be done? *Speak Out! 43*, 3-7.

Gilbert, J. B. (2012). *Clear speech. Pronunciation and listening comprehension in North American English* (4th ed.). Cambridge: Cambridge University Press.

Grant, L. (2014). Prologue to the myths: What teachers need to know. In L. Grant (Ed.), *Pronunciation myths: Applying second language research to classroom teaching* (pp. 1-33). Ann Arbor: University of Michigan Press.

Kuhn, T. S. (1970). *The structure of scientific revolution*. Chicago: University of Chicago Press.

Kuhn, T. S. (1977). *The essential tension: Selected studies in scientific tradition and change*. Chicago: University of Chicago Press.

Larsen-Freeman, D. (2003). *Teaching language: From grammar to grammaring*. Boston: Heinle.

Lewis, M. (1993).*The lexical approach: The state of ELT and a way forward*. Boston: Thomson Heinle.

Lewis, M. (1997). *Implementing the lexical approach*. Boston: Thomson Heinle.

Long, M. H., & Robinson, P. (1998). Focus on form: Theory, research and practice. In C. J. Doughty & J. Williams (Eds.), *Focus on form in classroom second language acquisition* (pp. 15-41). Cambridge: Cambridge University Press.

Munro, M. J. (2008). Foreign accent and speech intelligibility. In J. G. Hansen Edwards & M. L. Zampini (Eds.), *Phonology and second language acquisition* (pp. 193-219). Philadelphia: John Benjamins.

Nunan, D. (2004). *Task-based language teaching*. Cambridge: Cambridge University Press.

Piske, T., Flege J. E., MacKay, I. R. A., & Meador, D. (2011). Investigating native and non-native vowels produced in conversational speech. In M. Wrembel, M. Kul, & K. Dziubalska-Kołaczyk (Eds.), *Achievements and perspectives in the acquisition of second language speech: New sounds 2010: Vol. 2* (pp. 195-205). Bern: Peter Lang.

Post, B., & Nolan, F. (2011). Data collection for prosodic analysis of continuous speech and dialectal variation. In A. C. Cohn, C. Fougeron, & M. Huffman (Eds.), *The Oxford handbook of laboratory phonology* (pp. 538-547). Oxford: Oxford University Press.

Pritchard, A., & Woollard, J. (2010). *Psychology for the classroom: Constructivism and social learning*. Oxon: Routledge.

Ricard, E. (1986). Beyond fossilization: A course on strategies and techniques in pronunciation for advanced adult learners. *TESL Canada Journal,* s1, 243-253.

Strange, W., & Shafer, V. L. (2008). Speech perception in second language learners: The reeducation of selective perception. In J. G. Hansen Edwards & M. L. Zampini (Eds.), *Phonology and second language acquisition* (pp. 159-193). Philadelphia: John Benjamins.

Spada, N. (2011). Beyond form-focused instruction: Reflections on past, present and future research. *Language Teaching, 44*, 225-236.

Szczepek-Reed, B. (2011). *Analyzing conversation: An introduction to prosody*. Basingstoke: Palgrave Macmillan.

Warner, N. (2011). Methods for studying spontaneous speech. In A. C. Cohn, C. Fougeron, & M. Huffman (Eds.), *The Oxford handbook of laboratory phonology* (pp. 621-633). Oxford: Oxford University Press.

Watkins, M. A, Rauber, A. S., & Baptista, B. O. (Eds.). (2009). *Recent research in second language phonetics/phonology*. Cambridge: Cambridge Scholars.

Widdowson, H. G. (1984). *Explorations in applied linguistics 2*. Oxford: Oxford University Press.

Widdowson, H. G. (2000). On the limitations of linguistics applied. *Applied Linguistics, 21*(1), 3-25.

Zampini, M. L. (2008). L2 speech production research: Findings, issues, and advances. In J. G. Hansen Edwards & M. L. Zampini (Eds.), *Phonology and second language acquisition* (pp. 219-251). Philadelphia: John Benjamins.

APPENDIX A

Outline of the ConSpA

The idea for the connected speech-based approach (ConSpA) initially arose when I noticed in advanced/proficiency EFL classes I taught at the time that even such highly advanced learners were not able to comprehend authentic L1 English. I came to realize that there is a significant mismatch between students' use of rhythm and resulting connected speech phenomena, and respective use of such phonological features by L1 speakers. Interestingly, J. D. Brown (Brown & Kondo-Brown, 2006a; Brown, in press a) started his inquiry into teaching connected speech in a very similar way when a Chinese student once asked him: "Why is it that I can understand you in class, but cannot understand you when you talk to other American teachers?"

Connected speech is a phonological aspect that is virtually never taught in ESL/EFL classes because there are so many apparently random rules that, in isolation, bear no real meaning on their own, thus making connected speech phenomena practically impossible to teach by standards of current methodology. Further, rhythm and connected speech are rather intangible phonological aspects with very low perceptual saliency, which makes even noticing extremely difficult. The question, therefore, is how to make such phonological aspects processable and therefore teachable in a meaningful manner?

The solution proposed by the ConSpA is to present connected speech as consequence of the workings of English prosody. Prosody stands at the center of contemporary pronunciation pedagogy because of its meaning and context-generating quality and therefore offers a possibility to introduce connected speech in a meaningful context. As I have argued elsewhere (Euler, 2014a, p. 105-106), "connected speech becomes teachable and processable by helping students develop a solid understanding of prosodic areas like rhythm, thought grouping and primary stress allocation. In doing so the issue of individual rules, in turn, also becomes less of a problem because they will now appear as somewhat of a logical consequence." This results in the following kind of syllabus (see Euler, 2014a, 2014b for a detailed presentation), usable either in intensive courses or as a subsyllabus in regular ESL/EFL programs (Euler, 2014b):

Rhythm -> Intonation units -> Pitch & intonation contours -> Coalescent assimilation, deletion & reduction -> Linking & regressive assimilation <-> Sounds & positional variation

After students have explored rhythm and other prosodic aspects, they will see that something has to happen in sequences of unstressed function words in order to maintain stress-timing (see Euler, 2014c). By the time the prosody components are completed, they will have come across connected speech phenomena (with the teacher initiating some awareness-raising in such situations) so many times that they are, indeed, highly motivated to study this area systematically. The problem of individual rules can further be treated by "creating larger categories – coalescent assimilation, deletion and linking – and treating individual rules as possible instances that students can discover under the same framework (such as deletion)" (Euler, 2014a, p. 106).

Methodologically, in the course reported in this paper the Celce-Murcia model of communicative pronunciation teaching (Celce-Murcia et al., 2011) was predominantly used, which follows the following sequence:

Analysis -> Listening discrimination -> Controlled practice -> Guided practice -> Free practice

However, some task-based lessons were also included in which students start with processing input, during which they realize a need for language form which is then addressed at the end (e.g., Willis & Willis, 2007). In the ConSpA this is achieved through students not being able to understand L1 English and wanting to be able to decode the stream of speech (cf. Cauldwell, 2013). This is a cognitive window of opportunity for language practice. Once students realize that this is no "magic," that listening comprehension difficulties stem from not being able to recognize and process a distinctive set of phonological features, they will be very motivated to work on these features (see Euler, 2014c for a detailed discussion of task-based pronunciation teaching under a ConSpA framework). In addition, research has repeatedly shown that many learners do, in fact, strive for a native-like accent (e.g., Derwing, 2003) (as the students in the present study reported too), so these two aspects (comprehension and production) can easily be addressed complementing each other. This possibility is highly useful as it is very much in accordance with the focus on form system underlying much of structure teaching in task-based methodology (Doughty & Williams, 1998; Long & Robinson, 1998).

References

Brown, J. D. (in press a). *Shaping students' pronunciation: Teaching the connected speech of North American English*. Honolulu, HI: University of Hawaii at Manoa.

Derwing, T. M. (2003). What do ESL students say about their accents? *Canadian Modern Language Review*, *59*, 545-564.

Doughty, C. J., & Williams, J. (1998). Pedagogical choices in focus on form. In C. J. Doughty & J. Williams (Eds.), *Focus on form in classroom second language acquisition* (pp. 197-261). Cambridge: Cambridge University Press.

Willis, D., & Willis, J. (2007). *Doing task-based teaching*. Oxford: Oxford University Press.

APPENDIX B

Rating system[6]

Measure	Cambridge ESOL CEFR bands	Description
1	B1 Band 3 & B2 Band 1	Mostly intelligible. *Some control* of phonological features at word and utterance level.
2	B1 Band 4 & B2 Band 2	(between 1 and 3)
3	B1 Band 5 & B2 Band 3	Is intelligible. Intonation, stress, rhythm, & sounds *are often* accurate, but still some *clear* deviations.
4	B2 Band 4 & C1 Band 2	(between 3 and 5)
5	B2 Band 5 & C1 Band 3	Intonation, stress, rhythm, & sounds *are mostly* accurate; some deviations of *little bearing*.
6	C1 Band 4 & C2 Band 2	(between 5 and 7)
7	C1 Band 5 & C2 Band 3	Intonation, stress, rhythm, & sounds *are* accurate; *very few minor* deviations.
8	C2 Band 4	Almost native-like, but something does not seem to add up. (between 7 and 9)
9	C2 Band 5	Pronunciation is native-like

Measure: Continuation from 1 (not entirely intelligible) to 9 (native-like).
Cambridge/CEFR band: How the Measure value relates to the CEFR assessment bands.
Description: Descriptions are designed as broad landmarks. They are based on Cambridge's respective band descriptors but are modified as Cambridge English does not clearly differentiate between levels C1 and C2. The definitions are to be understood together with the continuation of the scale and the general CEFR level standards.
Compare Cambridge English's "Assessing speaking performance" sheets under:
https://www.teachers.cambridgeesol.org/ts/teachingresources/resourcedetails?resId=9016

[6] Each CEFR level is assessed in 5 bands, with Bands 1, 3 and 5 having a level description (the others being in-between values). All levels overlap in that, for example, Band 5 B2 equals Band 3 C1. Likewise, Band 1 B2 equals Band 3 B1. Band 3, therefore, represents its respective level and equals either the lowest band of the next higher level or the highest band of the next lower level.