SSLA, **31**, 609–639.
doi:10.1017/S0272263109990040

COMPREHENSION-BASED PRACTICE

The Development of L2 Pronunciation in a Listening and Reading Program

Pavel Trofimovich, Patsy M. Lightbown, Randall H. Halter, and Hyojin Song
Concordia University

We report the results of a 2-year longitudinal comparison of grade 3 and grade 4 English-as-a-second-language learners in an experimental, comprehension-based program and those in a regular (i.e., more typical) language learning program. The goal was to examine the extent to which sustained, long-term comprehension practice in both listening and reading—in the virtual absence of any speaking—can help develop learners' second language (L2) pronunciation. We analyzed learners' sentences from an elicited imitation task using several accuracy and fluency measures as well as listener ratings of accentedness, comprehensibility, and fluency. We found no differences between the two programs at the end of year 1. However, at the end of year 2, there were some differences—namely, in the listener ratings of fluency and comprehensibility—that favored learners in the

The research project on which this study is based involved the participation of many research assistants and colleagues who contributed to the original reports cited in the reference list. We are grateful for all the ways in which they made this study possible. Most of all, we are grateful to Allan Forsyth and William Francis Mackey for inviting us to observe and evaluate student language learning in this unusual program. It has had a profound impact on our understanding of language teaching and learning. We are also grateful to Erica Vukmanic and Kathryn MacFadden-Willard for their numerous contributions to our most recent analyses and to Sarita Kennedy and four anonymous *SSLA* reviewers for their valuable feedback on earlier drafts of this manuscript. This research was made possible through funding from the Ministère de l'Éducation du Nouveau-Brunswick (to Lightbown), with additional funding from the *Fonds québécois de la recherche sur la société et la culture* and the Social Science and Humanities Research Council of Canada (to Lightbown and Trofimovich), and support to the Centre for the Study of Learning and Performance.

Address correspondence to: Pavel Trofimovich, Concordia University, Department of Education, 1455 de Maisonneuve Blvd. West, Montréal, Québec, Canada H3G 1M8; e-mail: pavel.trofimovich@concordia.ca.

regular program. These findings highlight the beneficial effects of comprehension practice for the development of L2 pronunciation but also point to some potential limits of this practice.

In his 1917 book *The Scientific Study and Teaching of Languages* (republished in 1968), Palmer noted that learners' success in mastering a foreign language appears to be "in direct ratio to the degree in which they observe the natural laws of [that] language" (p. 46). Central to Palmer's statement is the idea that success in second language (L2) learning appears to depend on the exact nature and the extent of learners' experience with that language. This idea remains as pertinent to L2 learning and teaching now as it was almost a century ago. A great deal of research to date has been devoted to the study of just how learners should experience the language they are learning so that they can acquire it efficiently. Such research has focused, for example, on different types of language experiences, often discussed in terms of how rich the linguistic information (i.e., input) addressed to learners should be (Krashen, 1985; Lightbown, 1985), how learners perceive and internalize the input they receive (Long, 1991; Schmidt, 1990), or what particular kinds of language practices learners engage in (Long, 1981; Swain, 1985). According to Palmer, however, the learning experience that seems to be most beneficial is one he termed "passive work" (p. 48). This learning experience (which, we admit, is not passive at all) involves large amounts of listening and reading in a L2. The focus here is listening and reading practice, which will be discussed in relation to its role in the development of L2 pronunciation.

We report the results of a 2-year longitudinal comparison of two English-as-a-second-language (ESL) programs for grade 3 and grade 4 students in New Brunswick, Canada. One program was a typical ESL program in which students were mostly involved in speaking activities and a minimal amount of reading and writing practice. The other program was innovative in that it was based entirely on listening and reading activities. In this experimental program, students read stories and other English material and listened to accompanying audio recordings, independently, without lessons, tests, interaction with other students, or feedback from their teachers. Although such practice might be expected to develop students' comprehension of English, it is more difficult to predict the extent to which students might develop their ability to speak the L2 in the virtual absence of any speaking practice. To answer this question, we compared students in the two programs using several measures of pronunciation accuracy and fluency, with the goal of examining the role of comprehension practice in the development of L2 pronunciation.

COMPREHENSION-BASED APPROACHES TO L2 LEARNING

The idea that comprehension practice is central to L2 learning is clearly not a novel one. Over 500 years ago, for example, scholars and teachers of the Italian Renaissance stressed the importance of texts and reading in the learning of Latin and Greek and wrote of the significance of exposing learners to good models in these classical languages (Musumeci, 1997). However, it was not until the mid-1960s, with the advent of teaching approaches such as total physical response (Asher, 1965) and the natural approach (Krashen & Terrell, 1983), that comprehension-based teaching found its way into mainstream language classrooms in North America. Although these approaches make use of different pedagogical techniques (see Blair, 1982, for an overview), they are all based on a common assumption—namely, that experience in listening to L2 speech and reading L2 texts lays the foundation for language ability, including the ability to speak.

Krashen's (1985) input hypothesis and its subsequent extensions (e.g., Krashen, 2003) offer perhaps the most widely discussed theoretical justification for the importance of comprehension practice in L2 development (see Macnamara, 1973). In short, the input hypothesis holds that understanding messages (from oral and written input) is the only way in which humans acquire languages and that speaking is a consequence of language acquisition, not its cause. Speaking, which includes the ability to use the morphology and syntax of the L2 spontaneously and accurately, cannot be taught directly but emerges on its own. For Krashen, a good way to promote language acquisition is to expose learners to large amounts of comprehensible spoken language (Krashen & Terrell, 1983) and to engage learners in extensive reading activities (Krashen, 1993). Over the years, many premises that underlie this hypothesis, which include the claim that understanding messages is not only necessary but sufficient for learning a L2, have been challenged (e.g., White, 1987). What remains undisputed, however, is the basic idea that comprehension practice is beneficial for L2 learning.

BENEFITS OF COMPREHENSION-BASED PRACTICE

At least four strands of research support the beneficial role of comprehension practice in L2 development. The first line of evidence comes from studies of child L2 learning. It appears that children, when immersed in a L2 environment, often go through a silent period (Ervin-Tripp, 1974; Winitz, Gillespie, & Starcev, 1995). For example, Ervin-Tripp, who studied the acquisition of L2 French by 4- to 9-year-olds enrolled in French-language schools in Geneva, showed that many children "said nothing for many months" (p. 115), whereas others started speaking

6–8 weeks into their immersion in the classroom setting. The length of this silent period may vary from 1 to 3 months (Dulay, Burt, & Krashen, 1982) or may even extend up to 6 months (Winitz et al.). In a survey of 47 child L2 learners in Australia, Gibbons (1986) reported that the silent period varied between 0 days for some children and 56 days for others, with a mean of about 2 weeks. The precise nature of the learning that occurs during the silent period is unclear. However, extensive exposure to L2 input without any pressure to speak is probably beneficial in helping children build morphosyntactic and phonological knowledge. Adults may be less likely to profit from this kind of learning because they often must speak their L2 immediately for many social and economic reasons (Winitz et al.).

Other evidence of positive effects of comprehension practice comes from direct comparisons of individual tasks and of longer instructional interventions that include input or output components. Even a brief input experience (at least under certain conditions) can lead to improvements in some L2 skills. For example, in a comparison of two tasks, Izumi and Izumi (2004) showed that ESL learners were more accurate in tests of English relativization after completing a picture sequencing (i.e., comprehension) task than after a picture description (i.e., production) task (cf. De Jong, 2005). Learners' performance is also positively influenced by more focused comprehension experience. VanPatten and his colleagues have carried out numerous investigations to examine the effectiveness of one kind of comprehension practice (termed *processing instruction*) in relation to a more traditional language teaching, which emphasizes production of the L2. In these studies, processing instruction has often resulted in improved learner performance in both comprehension and production (e.g., VanPatten, 2004).

Positive effects of comprehension practice on production skills have additionally been shown in studies of content-based teaching (i.e., instruction that combines academic content with language learning objectives). Content-based instruction, particularly in a university setting, involves large amounts of aural and written input (from academic lectures and readings). This makes it possible to document how this input impacts not only learners' content knowledge but also their language ability, which includes production skills. Researchers at the bilingual (French-English) University of Ottawa have carried out numerous studies on the effectiveness of content-based instruction (e.g., Burger, Wesche, & Migneron, 1997). Learners in content-based classrooms master the subject matter and improve their L2 listening and reading skills (Hauptman, Wesche, & Ready, 1988). It is important to note that these learners also improve in L2 writing (Ready & Wesche, 1992) and speaking (Burger & Chrétien, 2001), which suggests that long-term comprehension practice has positive effects on the development of L2 production skills even when such skills are not emphasized in training.

Research in content-based instruction for children—such as immersion programs in Canada—has also shown that both comprehension and production skills develop in contexts in which the emphasis is on comprehension (Genesee, 1987).

Some of the most compelling evidence for the role of comprehension practice comes from studies of delayed speech production (e.g., Asher, 1969; Asher, Kusudo, & de la Torre, 1974; Winitz & Reeds, 1973). In one such study, Postovsky (1974) compared two groups of adult L2 learners of Russian enrolled in a 12-week course with an intensive oral practice component. The only difference between the two groups was that, for the first 4 weeks of the course, the learners in the experimental group did not speak in class. The results were striking: Six weeks into the course, the experimental group outperformed the control group on measures of speaking, reading, and writing. Although, by the end of the course, the test results for the two groups were similar, the experimental group still performed better than the control group in listening comprehension. Gary (1975) showed a similar advantage for child L2 learners of Spanish who did not speak for the first 14 weeks of a 22-week course. These learners outperformed the control group on measures of listening and performed similarly to the control group in speaking. Intensive comprehension experience without any pressure to speak (at least early on in learning) thus appears to accelerate the development of listening comprehension and may also have some benefits for speaking.

ROLE OF COMPREHENSION IN L2 PHONOLOGICAL LEARNING

Although comprehension practice has been found to promote both comprehension and production in the L2, its role in the development of L2 pronunciation is less well understood. Here, we will use the term *pronunciation* to refer to learners' ability to produce segmental and suprasegmental aspects of a L2 both accurately and fluently. One common view of the relationship between speech perception (i.e., comprehension) and speech production is that accurate perception is required for accurate production (Flege, 1995; Wode 1996). This perception-first view assumes that learners' perception (i.e., the ability to identify or discriminate L2 segments, stress patterns, or intonation contours accurately) often surpasses, and therefore precedes, their production (i.e., the ability to produce these aspects of L2 phonology accurately). However, failure to pronounce language accurately because of an inability to perceive particular features may be more typical of beginners (Baker, Trofimovich, Flege, Mack, & Halter, 2008), and accurate perception is no guarantee of accurate production. Advanced learners, whose perception and production abilities are more closely related, sometimes

perceive L2 segments more accurately than they produce them (Flege, Mackay, & Meador, 1999).

The perception-production link is most clearly seen in results of perceptual training studies. These studies show that teaching learners to identify or discriminate a particular feature of L2 speech not only improves their ability to perceive this feature but also facilitates their ability to produce it (Bradlow, Akahane-Yamada, Pisoni, & Tohkura, 1999; Bradlow, Pisoni, Akahane-Yamada, & Tohkura, 1997; Lambacher, Martens, Kakehi, Marasinghe, & Molholt, 2005). For example, Bradlow and colleagues (1997, 1999) have shown that Japanese learners of English who were trained over 3–4 weeks to identify English /r/ and /l/ (a difficult contrast for these learners) improved in their ability to produce these consonants and retained this learning gain even after a 3-month delay. A similar transfer of perception training into the production domain has been shown for phonologically delayed children (Rvachew, 1994), which suggests that perceptual bases of production are not specific to any one population of learners. Although there is some variability among learners in how their perception and production abilities are related, in that learners' production abilities may sometimes surpass their perception abilities (Sheldon & Strange, 1982), perception appears to be an important component of accurate production.

MOTIVATION FOR THE CURRENT STUDY

Most previous research on the relationship between perception and production has focused on investigating this relationship over a short term, using data gathered in a single session or relying on a brief training intervention (e.g., Bradlow et al., 1997, 1999). However, a small number of researchers have studied the perception-production link over a longer term by examining the effects of sustained perception experience on L2 pronunciation. Neufeld (1978) exposed 20 native speakers (NSs) of English to videotaped recordings of short utterances in Japanese and Chinese in 18 1-h lessons. The utterances were not glossed, and the meanings of words and sentences were never presented to the learners because the intention was solely for the learners to acquire nativelike pronunciation of each sentence. During the first 12 lessons, the learners listened to utterances and were, at times, asked to distinguish individual segments in words or to trace intonation contours for some statements. In the final six lessons, they imitated the utterances heard. To examine pronunciation, learner imitations from the final lesson were presented to six judges for rating. The judges rated almost half of the learners as NSs of Japanese or Chinese.

In another study, McCandless and Winitz (1986) compared the pronunciation of two groups of adult learners of German after a much

longer intervention than that described by Neufeld (1978). One group of learners attended an intensive comprehension-based course in which they engaged in meaningful activities (e.g., cooking meals and playing sports) and completed a large amount of listening homework (for a total of 240 h of instruction). Of the four instructors for this course, two were always present in class at any one time, modeling conversations, providing learners with task directions, and rephrasing speech in response to learner comprehension problems. The aim was for the learners to listen and understand German; therefore, speaking and imitation were discouraged. The other group of learners attended a more traditional course of German (for a total of 224 h of contact), which focused on oral practice, explicit explanations of grammar rules, and teacher feedback. At the end of the instruction period, the learners' imitations of five German sentences were presented to four NSs of German for rating. The comprehension group was rated as having a more authentic German accent than the traditional group, and both groups outperformed a control group (which received no instruction) but fell short of the ratings given to NSs of German.

Although these findings suggest that extensive perception experience might enable learners to pronounce the L2 in a nativelike manner, several questions remain unanswered. First, Neufeld's (1978) findings have little to do with comprehension practice, as the learners in his study received essentially form-focused perceptual training with no meaning attached to any input. The results reported by McCandless and Winitz (1986) are more revealing of the value of comprehension practice; however, their comprehension course involved highly structured, sequenced materials, homework, tests, and teacher feedback. These aspects of the course make it difficult to determine how L2 pronunciation develops in response to comprehension experience unaccompanied by syllabi, homework, assessment, and teacher intervention. Additionally, learners in McCandless and Winitz's program received little comprehension input from reading, as this skill was only introduced in the final 2 weeks of the course. Input from reading may enhance the impact of input from listening by providing visual support for auditory input. In both studies, pronunciation was evaluated through a limited range of measures, using only NS ratings of accentedness or nativelikeness. Finally, because both of these studies focused on the abilities of adults, it is uncertain whether these same effects would be observed in children.

There is no previous research that can answer the question of whether, and to what extent, sustained, long-term comprehension practice in both listening and reading (without structured classroom activities, oral interaction, teacher input, or tests and in the virtual absence of any language exposure outside the classroom) can help develop young learners' L2 pronunciation ability. Here, the objective was to answer these questions and address Asher's (1969) call, made 40 years ago, to

investigate "the amount of listening training which is necessary to produce a 'perceptual readiness' for speaking" (p. 17). To our knowledge, this is the first study to address Asher's call.

THE CURRENT STUDY

This study was conducted as part of a larger research project carried out in the mid-1980s and early 1990s to evaluate an experimental ESL program for young francophone learners of English in French-language elementary schools in New Brunswick, Canada (Lightbown, 1992a; Lightbown, Halter, White, & Horst, 2002). At the time of the study, students began learning ESL when they entered grade 3 (i.e., approximately 8 years old). The regular ESL program was a modified aural-oral program that involved students in question and answer activities, dialogues, songs, and a minimal amount of reading and writing practice. ESL classes were sometimes taught by an ESL specialist whose responsibility was to provide English-language teaching for most or all of the classes in the school. More typically, however, ESL classes were taught by regular classroom teachers, all L2 speakers of English, many of whom had no special training in ESL and some of whom expressed concern about whether they knew the language well enough to teach it.

In 1985, the Ministry of Education of New Brunswick developed an experimental ESL program that was implemented in three French-language school districts. Created by Forsyth and Mackey, the program was comprehension based: Students read stories and other English material and listened to accompanying tape recordings, independently, without lessons, tests, interaction with other students, or feedback from their teachers. From a theoretical perspective, the program recognized the important role of receptive skills in L2 development. From a pedagogical perspective, it provided students with materials appropriate for their level and allowed them to progress at their own pace. From a purely practical perspective, the program alleviated a shortage of specially trained teachers and of teachers fluent in English (Forsyth, 1990).[1]

In contrast to the regular program, the experimental program ESL classes took place in a specially equipped classroom that provided a cassette player and a headset for each student. Most classrooms looked like miniature language labs, with small carrels arranged in the middle of the room or around the perimeter. In each classroom, there was a large collection of books, and with each book or set of books, there was a cassette tape on which the content of the book(s) had been recorded. Some of the audio materials were commercial products purchased together with the books; others were recorded

specifically for the program. The materials included simple children's books with one line of text that accompanied a full-page picture, storybooks with more text per page, ESL readers in both straight text and comic book format, picture dictionaries, age-appropriate ESL textbooks, and works of nonfiction such as simple science books or biographies. In short, students had a very wide range of materials available to them.

At the beginning of class, students would select the materials they wished to read and listen to during the period. There were some constraints on their choice. Certain materials were considered appropriate for those just starting the program, whereas others were considered to be of greater difficulty; students were expected to use the easier materials for a period of time before going on to the more difficult ones. Nevertheless, there was a wide variety of materials at each level, which were not graded in any strict linguistic sense; the students were not guided through the materials in any fixed sequence. During each half-hour period, students worked entirely on their own and placed check marks in a simple log book to indicate which materials they had read. Thus, there was no teaching, no testing, no interaction, and no probing of students' comprehension. One year of this program involved approximately 90 h of exposure to listening and reading input (30 min a day, 5 days a week).

Students' performance in the experimental program was evaluated extensively and compared to the performance of students in the regular ESL program (Lightbown, 1992a; Lightbown et al., 2002); however, these evaluations have little to say about the development of students' L2 pronunciation. Therefore, here, we chose to revisit the original oral data from the regular and experimental programs (collected between 1986 and 1988) to explore how students' L2 pronunciation developed over time.

METHOD

Participants

The original participant sample included 99 francophone grade 3 students. These participants were drawn randomly from 20 intact ESL classes in four school districts in northern New Brunswick. (In one of these districts, schools implemented either the experimental or the regular program; in two of these districts, schools implemented only the experimental program; and in the fourth district, schools implemented only the regular program.) Of these 20 classes, 8 were regular ESL classes and 12 were in the experimental program. All students were followed for 2 years and were tested twice: at the end of grade 3 (year 1)

and grade 4 (year 2). Of the 74 students included in the final data analyses, 49 were in the experimental program (19 female, 30 male) and 25 were in the regular program (18 female, seven male). All of the students were taught by teachers for whom English was a L2. Twenty-five students were excluded from the final analysis because of (a) their absences for testing (n = 9), (b) the lack of testing in year 2 (n = 14), and (c) insufficient analyzable data (n = 2). All students were from Francophone families and had resided all their lives in francophone areas of New Brunswick. At the beginning of grade 3, these students were, on average, 8.4 years old (range: 7.7–9.6 years). Students had no prior instruction in English, and their knowledge of English was not considered when they were placed in their grade 3 classes.

Although all students seemed to be similar in most respects, several measures were administered to ensure that the students in the two programs did not differ in ways that could be expected to affect their learning of English at school. The first measure was a questionnaire addressed to parents. Using several Likert scales, parents rated their children's home language background (from –3 = French only to +3 = English only), their contact with English (from –14 = no contact with English to +14 = much contact with English), their interest in English and French readings (from –7 = no interest to +7 = great interest), as well as the extent to which the parents themselves were proficient in English (from –6 = limited proficiency to +6 = nativelike).[2] Although parents had some (albeit limited) knowledge of English (M = 0), for all students the home language was exclusively French (M = –2.9). All students had had little contact with English (M = –7.7) and were more interested in reading in French (M = 4.1) than in English (M = –3.2).

The second set of measures included an English aural vocabulary recognition (AVR) test, a French reading test, and the French version of the Otis-Lennon school ability test. The AVR test was administered to the students by the research team at the beginning of grade 3. This test required students to match pictures, drawn from eight domains (e.g., family, clothing, food), with isolated vocabulary items that they heard on tape. The Ministry of Education provided information about the students' performance on the remaining two measures: The French reading test (a province-wide French reading exam) had been administered to all students at the end of grade 1, and the Otis-Lennon test (a test of general academic ability) was administered at the end of grade 3. There were no statistically significant differences between the groups on any of these measures: AVR, $t(70) = 0.10$, $p = .92$; French reading, $t(70) = -0.46$, $p = .65$; Otis-Lennon, $t(70) = 0.54$, $p = .59$.[3] This suggests that the students in both programs were comparable at the outset of the study, at least with respect to their knowledge of and exposure to English and general academic ability. The characteristics of the two groups appear in Table 1.

Table 1. Background and language proficiency characteristics of participants by group

	Experimental		Regular	
Measure	*M*	*SD*	*M*	*SD*
Chronological age[a]	8.5	0.5	8.3	0.3
Home language background[b]	–2.9	0.4	–2.9	0.3
Contact with English[c]	–7.6	4.4	–7.8	4.3
Interest in French reading[d]	3.8	3.0	4.3	2.6
Interest in English reading[d]	–3.3	2.6	–3.1	3.5
Parental English ability[e]	–0.2	3.9	0.2	2.7
Grade 1 French reading	7.8	1.6	8.0	1.5
Grade 3 AVR pretest[f]	24.7	15.0	24.3	9.7
Otis-Lennon school ability	53.8	10.8	52.2	12.5

Note. [a]In years.
[b]Measured on a 7-point scale (–3 = French only, +3 = English only).
[c]Measured on a 29-point scale (–14 = no contact with English, +14 = much contact with English).
[d]Measured on a 15-point scale (–7 = no interest, +7 = great interest).
[e]Measured on a 13-point scale (–6 = limited proficiency, +6 = nativelike).
[f]Maximum possible score is 64.

Materials

The materials included six simple English sentences (five declarative and one interrogative). The sentences, which ranged between four and nine syllables in length, featured semantic content appropriate for 8- to 9-year-old beginning learners of English. The analyses assessed students' pronunciation accuracy on the basis of all attempted words in these sentences. However, two phonological or morphophonological features that are problematic for francophone learners were of particular interest: English /h/ (15 tokens) and possessive *–s* (which appeared as its /z/ allomorph, three tokens). The sentences and the distribution of their target phonological or morphophonological features are summarized in the Appendix.

First, for francophone learners of English, English /h/ is challenging in at least two ways. Even in advanced stages of learning, francophone learners tend to delete English /h/, a segment absent from the phonemic inventory of French (e.g., *My (h)at is red*, in which parentheses represent a deleted segment). Paradis and LaCharité (2001) reported that when Francophones use English loanwords in French (e.g., *hamburger*), /h/ is the only segment that is entirely deleted rather than replaced by a phonetically similar one. Second, even when Francophone learners develop the ability to produce English /h/ in appropriate contexts, they often start to epenthesize (i.e., insert) this segment in inappropriate contexts (e.g., *I hate [h]apples*, in which brackets represent an inserted

segment). This typically happens at the onset of a vowel-initial syllable (Janda & Auger, 1992). Therefore, analyzing the accuracy of English /h/ production by the students from the regular and experimental programs can measure how well they are able to learn a challenging aspect of L2 phonology. Across the six target sentences, there were seven word tokens that favored /h/-deletion and eight word tokens that favored /h/-epenthesis (see the Appendix).

Second, English possessive *–s* is also problematic for francophone learners of English. Along with phonologically similar third-person singular *–s*, the possessive *–s* appears to be acquired slowly and to varying degrees of mastery (e.g., Dulay & Burt, 1973). From the perspective of phonological learning, the three allomorphs of *–s* (i.e., /s/, /z/, and /əz/) may present learners with different degrees of difficulty. Wode (1980), for example, showed that /s/ emerged in the speech of German children learning English first, followed later by /z/ and /əz/. In fact, /z/ might appear to be more difficult to acquire than /s/ because the acoustic intensity of a voiced fricative is weaker than the acoustic intensity of its voiceless counterpart (Strevens, 1960), which makes /z/ less salient in the input and, as a result, less likely to be perceived. The allomorphs of *–s* (notably, /s/ and /z/) also appear to be harder to produce when the addition of this morpheme creates a consonant cluster (e.g., *Tim's* vs. *Mary's*), which increases the articulation difficulty for learners (Abraham, 1984). It appears, then, that the /z/ allomorph of the English possessive *–s* presents a challenge at both morphological and phonological levels. Analyzing the accuracy of students' production of English /z/ (albeit in a restricted phonetic context with relatively few tokens) could reveal their difficulties in learning a complex morphophonological aspect of the L2. Across the six target sentences, three word tokens elicited English possessive /z/ in a consonant cluster environment (see the Appendix).

Experimental Task

Students' oral expression was assessed by means of an elicited imitation task. In this task, participants heard auditory prompts that they repeated (one at a time) to the best of their ability. The premise of elicited imitation, a task used in both first language (L1) and L2 acquisition research, is that imitation is reconstructive in nature, which means that in order to repeat an utterance, listeners first need to decode it and then encode it again relying on their "productive linguistic system" (Slobin & Welsh, 1973, p. 496). Since its earliest uses (Fraser, Bellugi, & Brown, 1963), elicited imitation has emerged as a technique of choice that allows researchers to tap into learners' productive linguistic system at

the level of both morphosyntax (Erlam, 2006) and phonology (Trofimovich & Baker, 2006). It appears that elicited imitation yields results that are convergent with other measures of L2 performance (Gallimore & Tharp, 1981).

Despite these methodological strengths, elicited imitation may involve the obvious limitation that speakers simply mimic the utterance, including its phonological content. Speakers can more easily mimic shorter utterances, which are retrieved verbatim from short-term memory, than longer ones, which need to be reconstructed for production (Bley-Vroman & Chaudron, 1994). Here, we used sentences of varying lengths (four to nine syllables), comparable in complexity and length to those used previously with child NSs (McDade, Simpson, & Lamb, 1982). Although all sentences were short, our assumption was that sentences of different lengths would provide students with varying degrees of complexity and should yield an accurate estimate of their ability. We computed performance scores across all sentences to derive measures less dependent on any one sentence and its characteristics.

Speakers can also mimic an utterance more easily when they are able to repeat it immediately after hearing it. For example, McDade et al. (1982) showed that without at least a brief delay between an utterance and its repetition, child NSs can imitate sentences that they do not understand. Such utterances are available verbatim from short-term memory and require little processing to be produced. To estimate the likelihood of direct mimicry in the task reported here, we measured the time interval between each sentence and its repetition in a random selection of 15% of all recorded sentences (n = 132 sentences). There was a mean delay of about 1.64 s between the end of a sentence and the onset of its repetition. Because comparable delays have been observed in repetition tasks used with child NSs of similar age (Gathercole, Willis, Baddeley, & Emslie, 1994), it appeared unlikely that students were simply imitating sentences by rote.[4]

Procedure

Students were tested individually in a quiet location in their schools. The test, which was recorded on tape, was presented to each student over a headset connected to a tape recorder. Another recorder with a built-in microphone was used to record the entire testing session, including the students' productions. Students first listened to test instructions in French that were recorded on the same tape. They were told that they would hear a short story one sentence at a time and were asked to repeat each sentence exactly as it sounded, even if the sentence seemed too long to remember. Students were instructed to wait

for a short beep, recorded on the tape about 650 ms after each sentence, before speaking. Students then listened to a practice phrase and repeated it after the beep.

After the experimenter made sure that the student understood the instructions, the target sentences were played. Students heard each of the six target sentences in the same order (see the Appendix) by an interval of about 8 s. The sentences, recorded by a female NS of English, were spoken slowly and clearly (mean speech rate = 2.1 syllables per s [syll/s]) and sounded like a short story read aloud to young children. Upon hearing each sentence (e.g., *Oh no! Tom is eating Sam's hat*), students attempted to repeat it to the best of their ability (e.g., *Oh, no! Tom is eating Sam (h)at*; *Oh, no! Sam is eating Tom*; *Oh, no! Tom is [h]eating Sam hat*). Each student's entire testing session was later digitized, and individual sentences were saved separately and prepared for subsequent analyses (e.g., normalized for peak intensity and perceived loudness). All subsequent analyses were based on 888 sentences (74 students × two testing times × six sentences).

DATA ANALYSIS

Two sets of analyses were conducted. First, sentence-based measurements of speech accuracy and fluency were performed, which included students' repetition accuracy, pronunciation error ratios, and speech rate. Then students' sentences were presented to 20 NSs who rated these sentences for accentedness, comprehensibility, and fluency.

Sentence-Based Measurements

Repetition accuracy and pronunciation error ratio—the first two sentence-based measurements—were based on phonetic analyses of students' speech. These analyses were performed by a phonetically trained judge who listened to each sentence, transcribed it, and then compared each student's production of that sentence to the original sentence prompt. The judge calculated the number of words from each original sentence prompt that were correctly reproduced and computed the number of pronunciation errors in a student's repetition of each sentence prompt. To estimate the reliability of these judgments, we asked another rater (blind to the purposes of the study) to perform the same task using a random selection of 15% of all recorded sentences (n = 132 sentences; half from year 1, half from year 2). Intraclass correlations (ICCs; absolute agreement) for these two raters were very high:

ICC(2, 1) = .98 for judgments of words correctly reproduced and *ICC*(2, 1) = .93 for judgments of pronunciation errors.

The repetition accuracy score was computed by dividing the number of words correctly reproduced from the original sentence prompt by the total number of words in each sentence prompt. For example, if in response to the original sentence prompt *Oh no! Tom is eating Sam's hat* (seven words), a student said *Oh no! The Tom is eating Sam* (five original words repeated), this student's repetition accuracy score was 5/7 = .71. In calculating this score, a word was considered to be correctly repeated only if all its morphemes were correctly reproduced (e.g., *eating*, not *eat*; *Sam's*, not *Sam*) although the phonetic accuracy of word repetitions was disregarded (e.g., /a/ used in place of /æ/ in *Sam*). Word order and word insertion errors—that is, transposed or inappropriately inserted words (e.g., *Oh no! Is Sam is eating Tom*)—were also ignored; however, if a student repeated words or phrases from the original sentence prompt more than once (e.g., *(H)is dog name is Tom, Tom* or *His dog is name, the dog is name Tom*), each word's repetition was counted as correct only once. Repetition accuracy scores were first computed for each sentence (separately for year 1 and year 2 sentences) and then averaged for each student across the student's sentences to derive single final scores. These final repetition accuracy scores were used in all subsequent analyses.

The global measure of pronunciation errors was a score calculated by dividing the total number of words that contained a pronunciation error by the total number of words produced by the student in repeating a given sentence. Pronunciation errors were defined as omitted or inserted phonemes or morphemes, any phonemic vowel or consonant substitution errors, or errors of consonant cluster simplification. Examples of these included phonological or morphophonological errors commonly found in the speech of francophone learners of English (e.g., production of /ð/ as /d/, vowel substitution errors, /h/-deletion and /h/-epenthesis, lack of aspiration of voiceless stops in stressed syllables, omission of a final /z/ morpheme in a consonant cluster). Errors in sentence structure (including word order), morphology, or syntax as well as errors related to sentence prosody were ignored. In calculating errors, we considered only intelligible words so that we could clearly match the content of all original sentences with the students' repetitions. For example, if in response to the original sentence prompt *His hair is blue and his nose is red*, a student said, *(H)is (h)air is blue ? his nose is red* (in which the question mark represents an unintelligible word), this student's pronunciation error score was 2/8 = .25. In other words, for this student, out of eight words produced, two (*his* and *hair*) contained a pronunciation error. We opted to have a conservative measure of pronunciation accuracy and, thus, decided not to inflate pronunciation error rates by counting multiple pronunciation errors

per word. This controlled for extreme cases of variability in individual students' error counts. Some individual students produced certain words with a single error, whereas others made multiple errors. For example, if a student produced *Sam's* as [sam]—that is, omitting the /z/ morpheme and substituting /a/ for /æ/—only one error was recorded for this word.

Although the global pronunciation error score included all pronunciation errors, separate pronunciation error scores were calculated for the two target features (i.e., English /h/ and the /z/ allomorph of the English possessive *–s*). The error score for English /h/ was computed by dividing the total number of /h/ errors (both deletion and epenthesis) by the total number of /h/ words attempted by each student. For example, if a student produced *Oh no! Tom [h]eating (h)at*, this student's /h/ error score was 2/3 = .67. In other words, for this sentence, out of three words favoring /h/-deletion or /h/-epenthesis (i.e., *oh, eating*, and *hat*), the student made an error in two of them (producing *oh* correctly without epenthesis, but saying *eating* as *heating* and *hat* as *at*). Similarly, the error score for English /z/ was calculated by dividing the total number of /z/ errors (omissions, in this case) by the total number of words produced that contained English possessives. As with the repetition accuracy scores, pronunciation error scores were calculated for each student based on each sentence (separately for year 1 and year 2 sentences); these scores were then averaged across the six sentences to derive single final scores for each student. These final scores (global error score, /h/ error score, and /z/ error score) were used in all subsequent analyses.

The final sentence-based measurement was speech rate—an index of articulation fluency. Following previous investigations (e.g., Trofimovich & Baker, 2006), the speech rate was computed for each sentence by dividing the number of uttered syllables by the total duration of an entire utterance, including pauses. Syllable counts and duration measurements were carried out by the same phonetically trained judge who performed the repetition accuracy and pronunciation error counts. Durations were measured to the nearest millisecond by means of *Cool Edit 2000* digital speech-analysis software (Johnston, 1999). For example, if a student produced *His dog is name, the dog is name Tom*, this student's syllable count was 9. With a 5.336-s total duration of the utterance, the student's speech rate ratio (syll/s) was 1.69. To determine the reliability of these measurements, the blind independent rater who had earlier scored repetition accuracy and pronunciation errors was asked to perform syllable counts and duration measurements for a randomly chosen sample of 15% of all recorded sentences (n = 132 sentences). ICCs (absolute agreement) for the two raters were again very high: $ICC(2, 1)$ = .98 for syllable counts and $ICC(2, 1)$ = .99 for duration measurements. Speech rate ratios were first calculated for each sentence

(separately for year 1 and year 2 sentences) and then averaged across each student's six sentences to derive a single final speech rate ratio used in all subsequent analyses.

Listener Ratings

There were three listener-rated measures of students' oral performance: accentedness, comprehensibility, and fluency. *Accentedness* denotes the listeners' judgment of how closely learners' pronunciation of an utterance approaches that of a NS (Munro & Derwing, 1999). *Comprehensibility* refers to the listeners' perceptions of how easily they understand an utterance (Munro & Derwing). *Fluency* is defined as the listeners' judgment of how fluid an utterance sounds, spoken without undue pauses, filled pauses, hesitations, or dysfluencies such as false starts and repetitions (Trofimovich & Baker, 2006).

To obtain listener judgments, all recorded sentences were presented to 20 listeners for rating. Recruited from students in linguistics or English language teaching in two English-medium universities in Montreal, the listeners were, on average, 31 years old (range: 19–49). Nineteen listeners claimed English as the only language learned from birth; one listed both English and French as mother tongues. All listeners were proficient in at least one language other than English (e.g., French, German), and, as residents of Montreal, all had experience with French-accented English. Because rating the entire sample of 888 sentences would have been fatiguing for an individual listener, the sentences were randomly divided into two equal sets, with the six sentences from years 1 and 2 equally represented in each set. The 20 listeners were then randomly assigned to two groups of 10 listeners, and each group evaluated one set of sentences. One shortcoming of this arrangement, however, is that the two listener groups would never rate the same sentences, and it would not have allowed us to compute interrater reliability across all listeners. Thus, 120 individual recordings (20 for each sentence, drawn equally from years 1 and 2) were selected for rating by both listener groups. Listener ratings of these 120 common recordings were used to estimate interrater reliability.

All listeners made their judgments individually in a quiet room using a Macintosh G4 iMac computer and a Sony MDR-CD60 headset. The 384 sentences (64 students × six sentences) unique to each listener group and the 120 sentences common to both groups (20 students × six sentences) were presented in six randomized blocks (for a total of 504 sentences) using one of two random orders. Each block contained all productions of a specific sentence. Listeners were told which sentence the students were attempting to say so that they could compare

what they heard to their expectations of what the sentence should sound like. Listeners played back each sentence one at a time and rated it on three 9-point Likert scales on a response sheet: accentedness (from 1 = heavily accented to 9 = not accented at all), comprehensibility (from 1 = hard to understand to 9 = easy to understand), fluency (from 1 = not fluent at all to 9 = very fluent). Listeners were told to use the entire scale and were allowed to listen to each sentence as many times as they wished. The rating session lasted between 2.5 and 3 h.[5]

The ICCs (absolute agreement) computed for the 20 listeners' ratings of the 120 common recordings yielded a range of very high indexes: *ICC*(3, 20) = .86–.93 for accentedness, *ICC*(3, 20) = .90–.96 for comprehensibility, and *ICC*(3, 20) = .84–.96 for fluency. Given such a high rating consistency among the listeners, accentedness, comprehensibility, and fluency scores were calculated for each sentence by averaging across the 10 listeners' ratings in each listener group. As with the repetition accuracy and pronunciation error scores, this calculation was done separately for the year 1 and the year 2 sentences. Final scores of accentedness, comprehensibility, and fluency were then derived for each student by computing the mean for this student's six sentences. These final accentedness, comprehensibility, and fluency scores were used in all subsequent analyses.

RESULTS

For all statistical tests, the alpha level for significance was set at .05. The effect sizes are partial eta squared (η_p^2). A Bonferroni procedure was applied to adjust the level of significance for all tests of simple main effects (adjusted $\alpha = .0125$).

Sentence-Based Measurements

Repetition accuracy scores, which provided a baseline measure of accuracy, were analyzed first. Overall, students repeated approximately 60–70% of the words accurately, with individual scores ranging between 22% and 97% correct. To examine whether there were any between-group differences, the repetition accuracy scores were submitted to a two-way repeated measures ANOVA with group (experimental and regular) as a between-subjects factor and time (year 1 and year 2) as a within-subjects factor. This analysis yielded a significant main effect for time, $F(1, 72) = 8.60$, $p < .005$, $\eta_p^2 = .11$, with no significant effect for group, $F(1, 72) = 2.68$, $p = .11$, $\eta_p^2 = .04$, nor a

significant interaction, $F(1, 72) = 3.67$, $p = .06$, $\eta_p^2 = .05$. This finding suggests that both groups made an improvement (albeit small) in their repetition accuracy between year 1 and year 2. Although there was no statistically significant difference between the groups, the regular group demonstrated a slightly larger accuracy gain than the experimental group did. Descriptive statistics for both groups appear in Table 2.

In terms of global pronunciation accuracy, students made pronunciation errors in about 25% of all the words they produced, with individual error rates ranging between 0% and 61%. Errors specific to English /h/ and English possessive /z/ were far more prevalent: /z/ was omitted in about 40–60% of all words that contain the English possessive and /h/ was erroneous (deleted or epenthesized) in roughly 70–80% of all the /h/ words attempted by students. Deletion was, by far, the most predominant of all /h/ errors observed. In fact, /h/-epenthesis accounted for only 5% of all /h/ errors.

The three sets of error scores were submitted to separate 2 × 2 (Group × Time) repeated measures ANOVAs. The analysis of global error scores

Table 2. Descriptive statistics for all sentence-based measurements by group

	Experimental		Regular	
Measure	Year 1	Year 2	Year 1	Year 2
Repetition accuracy				
M	0.59	0.61	0.61	0.68
SD	0.13	0.13	0.13	0.10
Range	0.22–0.97	0.33–0.95	0.32–0.83	0.47–0.87
Pronunciation error score				
M	0.24	0.25	0.24	0.27
SD	0.08	0.10	0.09	0.09
Range	0.03–0.39	0.00–0.61	0.09–0.42	0.07–0.42
English /h/ error score				
M	0.72	0.79	0.71	0.71
SD	0.28	0.26	0.33	0.32
Range	0.00–1.00	0.00–1.00	0.00–1.00	0.00–1.00
English /z/ error score				
M	0.56	0.48	0.43	0.59
SD	0.27	0.29	0.30	0.35
Range	0.00–1.00	0.00–1.00	0.00–1.00	0.00–1.00
Speech rate (syll/s)				
M	1.89	1.90	1.86	1.92
SD	0.21	0.18	0.17	0.23
Range	1.35–2.30	1.54–2.28	1.54–2.13	1.49–2.29

yielded no significant main effect for time, $F(1, 72) = 3.38$, $p = .07$, $\eta_p^2 = .05$, or for group, $F(1, 72) = 0.43$, $p = .52$, $\eta_p^2 = .01$, and no significant interaction, $F(1, 72) = 0.52$, $p = .48$, $\eta_p^2 = .01$. The analysis of /h/ error scores similarly yielded no significant main effect for time, $F(1, 72) = 0.93$, $p = .34$, $\eta_p^2 = .01$, or for group, $F(1, 72) = 0.49$, $p = .49$, $\eta_p^2 = .01$, and no significant interaction, $F(1, 72) = 0.99$, $p = .32$, $\eta_p^2 = .01$. These results suggest that the two groups' global and /h/ error rates were comparable in year 1 and year 2. The analysis of /z/ error scores yielded only a significant interaction, $F(1, 72) = 8.32$, $p < .005$, $\eta_p^2 = .11$, with no significant effects for time, $F(1, 72) = 0.88$, $p = .35$, $\eta_p^2 = .01$, or for group, $F(1, 72) = 0.03$, $p = .86$, $\eta_p^2 = .001$. However, tests of simple main effects, conducted to explore this interaction, failed to locate any significant differences after a Bonferroni correction was applied. These results suggest that both groups were comparable in terms of their pronunciation error rates in year 1 and year 2 (see Table 2).

In the final set of analyses, speech rate ratios were analyzed using a similar 2 × 2 (Group × Time) repeated measures ANOVA. This analysis yielded no significant main effect for time, $F(1, 72) = 1.20$, $p = .28$, $\eta_p^2 = .02$, or for group, $F(1, 72) = 0.01$, $p = .94$, $\eta_p^2 = .001$, and no significant interaction, $F(1, 72) = 0.59$, $p = .45$, $\eta_p^2 = .01$. This finding suggests that the two groups produced English sentences at about the same speech rate in year 1 and year 2. In sum, analyses of sentence-based measurements revealed that the students in both the experimental and the regular group made comparable (albeit modest) improvement in their word repetition accuracy between year 1 and year 2 and that they did not differ (in either year 1 or year 2) in terms of how many pronunciation errors they made or how quickly they spoke.

Listener Ratings

A 2 × 2 (Group × Time) repeated measures ANOVA that examined accentedness scores yielded a significant main effect for time, $F(1, 72) = 55.18$, $p < .0001$, $\eta_p^2 = .43$, and a significant interaction, $F(1, 72) = 5.38$, $p < .05$, $\eta_p^2 = .07$, but no significant effect for group, $F(1, 72) = 1.59$, $p = .21$, $\eta_p^2 = .02$. Tests of simple main effects, conducted to explore the significant interaction, revealed that both groups were perceived as being less accented in year 2 than in year 1 (experimental group: year 1, $M = 4.08$; year 2, $M = 4.38$; regular group: year 1, $M = 4.16$; year 2, $M = 4.74$; $p < .0001$) and that the regular group was perceived as being less accented than the experimental group, but only in year 2 ($M = 4.74$ vs. $M = 4.38$; $p < .05$). This last difference, however, missed statistical significance after a Bonferroni correction. Mean accentedness ratings are plotted in Figure 1.

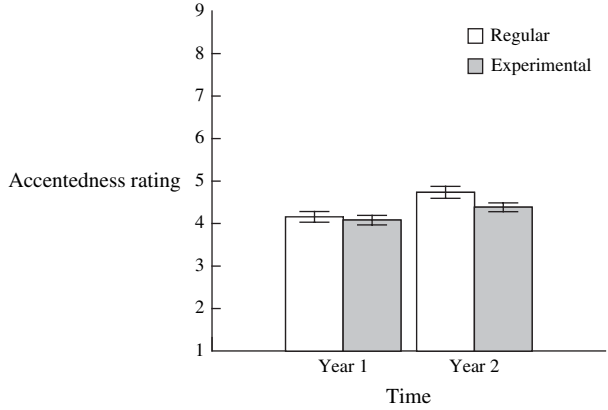

Figure 1. Mean accentedness ratings for regular and experimental group students in year 1 and year 2. Higher values represent less accented sentences. Brackets enclose ± 1 standard error.

A 2 × 2 (Group × Time) repeated measures ANOVA that examined comprehensibility scores yielded a significant main effect for time, $F(1, 72) = 97.99$, $p < .0001$, $\eta_p^2 = .58$, and a significant interaction, $F(1, 72) = 6.55$, $p < .05$, $\eta_p^2 = .08$, but no significant effect for group, $F(1, 72) = 2.90$, $p = .09$, $\eta_p^2 = .04$. Tests of simple main effects showed that both groups received higher comprehensibility ratings in year 2 than in year 1 (experimental group: year 1, $M = 4.32$; year 2, $M = 5.04$; regular group: year 1, $M = 4.52$; year 2, $M = 5.74$; $p < .0001$) and that the regular group received higher comprehensibility ratings than the experimental group, but only in year 2 ($M = 5.74$ vs. $M = 5.04$; $p < .05$). Mean comprehensibility ratings are plotted in Figure 2.

Finally, a 2 × 2 (Group × Time) repeated measures ANOVA that examined fluency scores revealed the same pattern of findings: a significant main effect for time, $F(1, 72) = 56.49$, $p < .0001$, $\eta_p^2 = .44$, and a significant interaction, $F(1, 72) = 4.24$, $p < .05$, $\eta_p^2 = .06$, with no significant effect for group, $F(1, 72) = 3.43$, $p = .07$, $\eta_p^2 = .05$. Tests of simple main effects further revealed that both groups received higher fluency ratings in year 2 than in year 1 (experimental group: year 1, $M = 4.96$; year 2, $M = 5.46$; regular group: year 1, $M = 5.12$; year 2, $M = 5.99$; $p < .0001$) and that in year 2 the regular group received higher fluency ratings than the experimental group ($M = 5.99$ vs. $M = 5.46$; $p < .05$). Mean fluency ratings are plotted in Figure 3.

In sum, analyses of listener ratings showed that the students in both groups were perceived as being less accented, more comprehensible, and more fluent in year 2 than in year 1. Additionally, in year 2, the students in the regular group were perceived as being slightly more comprehensible and fluent (but not less accented) than the students in the experimental group.

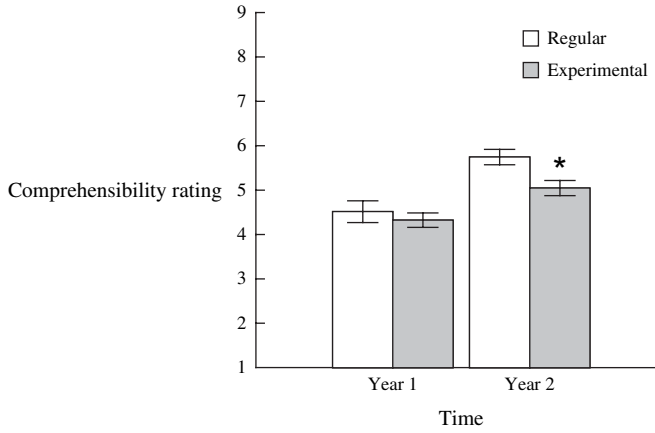

Figure 2. Mean comprehensibility ratings for regular and experimental group students in year 1 and year 2. Higher values represent more comprehensible sentences. Brackets enclose ± 1 standard error. An asterisk shows a statistically significant difference between the two groups (p = .012).

DISCUSSION

The goal of this study was to determine whether, and to what extent, sustained, long-term comprehension practice in both listening and reading helps develop L2 pronunciation ability. To this end, longitudinal

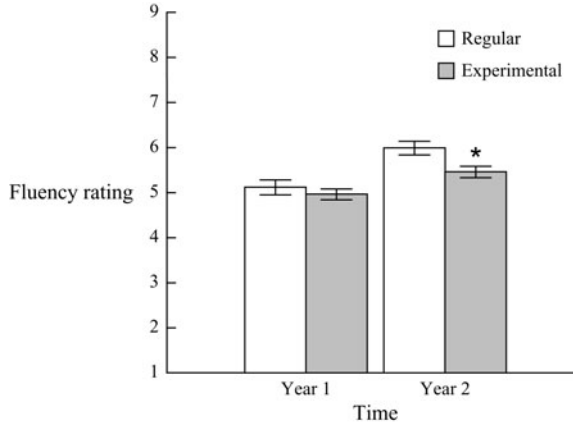

Figure 3. Mean fluency ratings for regular and experimental group students in year 1 and year 2. Higher values represent more fluent sentences. Brackets enclose ± 1 standard error. An asterisk shows a statistically significant difference between the two groups (p = .011).

comparisons of pronunciation accuracy and fluency were carried out for students in an experimental comprehension-based program (which involved listening and reading practice only) and in a regular program (which involved both aural and oral practice in addition to minimal amounts of reading and writing).

The analyses revealed no major differences between the regular and the experimental programs. The students in both programs made comparable (although modest) improvement in their repetition accuracy and, according to listener judgments, became less accented, more comprehensible (i.e., easier to understand), and more fluent in their speech between the first and the second year. It appears, then, that both types of learning experiences were beneficial for the development of students' pronunciation skills. The comparison of students' pronunciation errors (using global and specific accuracy measures) as well as of students' fluency (using a measure of speech rate) again yielded no differences between the regular and the experimental programs. Additionally, there was no improvement for either group over time. Accurate production of individual segments and syllables (including /h/ and possessive /z/, which are especially hard for Francophones) proved to be equally problematic for learners in both groups. The only group differences emerged in listener ratings of comprehensibility and fluency. At the end of the second year, students in the regular program were perceived as being more fluent and easier to understand than students in the experimental program, which reveals some potential limits of comprehension-based learning.

Perceptual Readiness for Speaking

At first glance, the finding that the students in the experimental program performed as well as the students in the regular program appears unremarkable. This result appears to be reminiscent of the outcomes of so-called method comparison studies in which, for example, audiolingual instruction was contrasted with cognitive code teaching (e.g., Chastain, 1969). These studies often yielded a finding of no difference, which led researchers to conclude that their results could not clearly distinguish between the methods being contrasted (see Long, 1980). However, our aim in contrasting the two programs was not motivated by a desire to compare teaching methods. In fact, the instruction in the experimental program was not based on any particular teaching method. Apart from an emphasis on comprehension through the materials made available to learners, it was the learners themselves who determined the course of their instruction, which took place in a low-stress context, with the freedom to choose what they found most interesting and with an implicit focus on the development of autonomous learning skills.

The regular group, which received classroom instruction, had access to many seemingly advantageous aspects of classroom teaching, including a structured syllabus, trained teachers, teacher feedback, output practice, and interaction. These aspects would seem to favor learning outcomes in the instructed setting versus the uninstructed comprehension-based context. Despite this apparent disadvantage for the students in the comprehension-based program, they performed just as well as the instructed regular group on nearly every measure. Seen from this vantage point, then, the finding of no difference between the two programs on most measures of pronunciation quality seems quite remarkable (see Gary, 1975, for a similar argument). The possible cause for this finding is the substantial difference between the two programs in the quantity (and perhaps the quality) of input, independent of the method of instruction (Lightbown, 1992b). Indeed, students in the experimental program received a great deal more input from reading and listening than students in the regular program. This input helped sustain the pronunciation ability of the experimental group, for at least 1 year, in the virtual absence of any speaking practice and exposure to English outside the classroom.[6]

However, at the end of the second year of their comprehension experience (after roughly 180 h of input), the students in the experimental program were judged by English listeners as being slightly harder to understand and less fluent than the students in the regular program. It appears that beyond the first year of comprehension-based practice, additional input from listening and reading was less effective in promoting fluent and clearly articulated L2 speech. The students in the experimental program were not speaking any more slowly than those in the regular program (as the groups did not differ in speech rate), yet they sounded significantly more dysfluent, which suggests that their speech featured additional content but was fraught with more pauses, hesitations, and false starts.[7] These lower ratings likely reflect the students' lack of direct practice in oral production—the very skill on which they were evaluated. These ratings also show possible limits on how long perceptual experience alone can sustain the development of accurate pronunciation.

It is possible that about 90 h of comprehension experience (i.e., 1 year of the program) represents, in the context of this study, what Asher (1969) termed the amount of perceptual experience needed for a perceptual readiness for speaking. If this assumption is valid, the students in the experimental program were ready to start speaking after about 1 year. Supplementing comprehension practice with oral production activities in the second year of the program might have given these students additional skills for translating their rich perceptual experience into accurate, fluent, and effortless L2 production (see DeKeyser, 2007, for a view of skill acquisition compatible with this idea).

Benefits of Listening and Reading for L2 Phonological Development

The results of this study raise an interesting question about the precise contribution of listening and reading practice to the development of L2 pronunciation. This question is addressed in theoretical frameworks that hold that knowledge of language structure (including phonology) emerges from language users' experience with language, particularly their experience with the lexicon (Bates & Goodman, 1997; Pierrehumbert, 2003). In essence, language users become attuned to the regularities they perceive in the language around them and create generalizations on the basis of these regularities. For example, NSs of English learn to expect that no real English word ends in /h/ because, in their experience with English sound structure, the probability of such a word occurring is close to zero. In fact, many researchers have now shown that language users, including L2 learners, are sensitive to the frequency of phonological, morphological, and syntactic regularities in the input (e.g., Goldschneider & DeKeyser, 2001), and that the structure of language users' linguistic knowledge closely corresponds to the properties of the input they receive (Ellis, 2007; Robinson & Ellis, 2008).

Through reading and listening to a great variety of authentic texts, the students in the comprehension-based program were exposed to very rich linguistic input, replete with varied, yet recurring, context-appropriate language. It is possible that this abundant lexical experience created opportunities for the students to engage in the kind of learning driven by language input. Early in this learning process, for example, the students would first notice and learn some frequent individual words (e.g., *toy*) and word co-occurrences (e.g., *Jane's toy*). After multiple opportunities to experience these words in different listening and reading contexts, the students would note regularities among them, noticing, for instance, similarities in phonology and morphosyntax. The students would then start to create a set of generalizations on the basis of these regularities. At the level of phonology, the students could infer that English /t/ at the onset of a stressed syllable is aspirated (e.g., *Tom* and *toy*), and at the level of morphosyntax, they could notice the structural configuration of the English possessive (e.g., *Tom's toy* and *Jane's dog*).

The experimental comprehension-based program is also an example of what many researchers believe to be a beneficial, if not ideal, context for L2 phonological learning (Bradlow et al., 1997; Trofimovich, 2008). In this program, students were exposed to spoken language that was highly variable in nature. The listening texts available to students were recorded by different individuals, represented different text genres, and featured a variety of speaking styles, rates, and intonation patterns. Variable spoken input is important for L2 phonological development

because it allows learners to focus on important differences in the input (e.g., phonemic distinctions) while learning to disregard other types of linguistically less important differences (e.g., allophonic variation and speaker-specific differences in speaking styles, rates, and pitch). Exposing learners to highly variable spoken input may thus help avoid learning that is specific to an individual speaker (i.e., the teacher) and an individual learning situation (i.e., the classroom)—the type of learning typical of many classrooms.

It may also be noteworthy that students in the experimental program received aural input only from the NSs whose voices were recorded on the tapes. These students, therefore, received only high-quality input in terms of the accuracy of the language heard and read. By contrast, students in the regular program also heard the speech of their peers. In the context of this (audiolingual) program, there was little spontaneous language use, which may have limited the type and number of pronunciation errors to which students were exposed. Nevertheless, if learners acquire their L2 abilities at least in part on the basis of input frequencies, exposure to imperfect pronunciation models might be expected to contribute to their own perceptions of how words and sounds are pronounced. For example, hearing imperfect pronunciation models could reinforce some common pronunciation errors and thus lead students to persist in perceiving and producing certain aspects of their L2 inaccurately.

In addition to receiving spoken input, students in the experimental program also had input from reading. For each class, the students selected texts that were of interest to them and listened to the accompanying audio recordings. There was a great deal of variability in how the students approached this task. Some students just listened without looking at the text. Other students traced the words with their finger while they listened. Others mumbled along with the audio track and followed the printed text, which raises the interesting possibility that comprehension practice through listening and reading was also successful at providing at least some students with opportunities for oral production. Without exception, however, all students had both print and speech available to them simultaneously, at all times.

One of the obvious advantages of having reading and listening input presented simultaneously is that printed text provides visual support for auditory input. Seeing a word in print while simultaneously hearing its pronunciation may help readers match a spoken word (e.g., /kæt/) with its orthographic shape (e.g., *cat*), a task notoriously hard for many beginning readers (Rayner, 1988). Hearing the spoken version of printed text also helps readers do the opposite task—convert printed text to speech. Some readers, especially beginners, often find themselves covertly sounding out the words they are reading. This process of subvocalization appears to be beneficial for the comprehension of texts (Daneman & Newson, 1992) and for memory storage and integration of ideas in a

text (Baddeley, Eldridge, & Lewis, 1981). By listening and reading at the same time, students in the experimental program had ample opportunities to notice how the written word relates to speech.

CONCLUDING REMARKS

In this study, we have shown that the students in the experimental program generally succeeded in being able to sound as accurate and just about as fluent and easy to understand as those whose learning was guided by a teacher in a more traditional program. However, we were unable to identify precisely which aspects of the experimental program contributed to this learning: emphasis on comprehension before production, large quantities of engaging listening and reading materials, high-quality input (in terms of acoustic quality and the accuracy of the language heard and read), low classroom anxiety, or a focus on independence and autonomy. Future research on comprehension-based learning may need to examine individual and combined contributions of these and other factors to learner success in L2 pronunciation development.

These conclusions, of course, are tentative and should be investigated further using other tasks and pronunciation measures. Nevertheless, what remains undisputed is the general value of comprehension practice for L2 learning. The comprehension-based program focused on here represents one of many possible approaches for enriching L2 teaching with comprehension practice through listening and reading. This program was student-centered: It encouraged students to take responsibility for their own learning and catered to individual differences in learning rate and interest. There was no pressure for students to perform. There was also no interaction and, consequently, no reinforcement of incorrect language forms from the speech of others. Additionally, the program provided students with high-quality input in a context in which it was often difficult to find teachers who were proficient in the L2 or trained to teach it. Implementing such a program, in its entirety, on a large scale, may, admittedly, prove difficult (however, see Di Biase, 1994, who described a similar program for teaching Italian in Australia). Nevertheless, many aspects of this program can offer a useful, if not absolutely essential, supplement to any language instruction, which would expand learners' opportunities for increased autonomy and enriched language input.

(*Received 8 December 2008*)

NOTES

1. The experimental program was being offered alongside the regular program until at least 1992, when the last research project that compared the two programs ended.

Precise information about the experimental program after 1992 was unavailable to the authors.

2. These Likert-type scales were derived from several individual questions (for more information, see Lightbown, 1992a).

3. Pretest data were not available for 2 of the 74 students.

4. As pointed out by an anonymous *SSLA* reviewer, it is impossible to rule out entirely the possibility that, even with a considerable delay between each sentence and its repetition, the students were subvocally rehearsing the sentences without processing them for meaning.

5. Although the listening session was relatively long, its length was comparable to listening sessions administered in previous studies of L2 speech learning (e.g., Derwing, Munro, & Wiebe, 1998). Possible listener fatigue effects were controlled for by presenting sentences to listeners in several randomized orders.

6. Posttest questionnaires revealed that, at the end of year 2, the students were comparable in terms of their contact with English outside the classroom. In fact, the students' contact with English outside the classroom was rated slightly lower at the end of year 2 (–7.6 and –8.8 for the regular and experimental groups, respectively) than at the beginning of year 1 (see Table 1).

7. An anonymous *SSLA* reviewer raised the interesting possibility that the obtained difference between the two groups in fluency rating (in the absence of a difference in speech rate) might be due to different patterns of pause placement by the two student groups. We plan to explore this possibility in future comparisons of suprasegmental features in the speech of both groups.

REFERENCES

Abraham, R. G. (1984). Patterns in the use of the present tense third person singular -s by university-level ESL speakers. *TESOL Quarterly*, *18*, 55–69.

Asher, J. J. (1965). The strategy of the total physical response: An application to learning Russian. *International Review of Applied Linguistics*, *3*, 291–300.

Asher, J. J. (1969). The total physical response approach to second language learning. *Modern Language Journal*, *53*, 3–17.

Asher, J. J., Kusudo, J. A., & de la Torre, R. (1974). Learning a second language through commands: The second field test. *Modern Language Journal*, *58*, 24–32.

Baddeley, A. D., Eldridge, M., & Lewis, V. (1981). The role of subvocalization in reading. *Quarterly Journal of Experimental Psychology*, *33*, 439–454.

Baker, W., Trofimovich, P., Flege, J. E., Mack, M., & Halter, R. (2008). Child-adult differences in second-language phonological learning: The role of cross-language similarity. *Language and Speech*, *51*, 316–341.

Bates, E., & Goodman, J. C. (1997). On the inseparability of grammar and the lexicon: Evidence from acquisition, aphasia, and real-time processing. *Language and Cognitive Processes*, *12*, 507–584.

Blair, R. W. (1982). *Innovative approaches to language teaching*. Rowley, MA: Newbury House.

Bley-Vroman, R., & Chaudron, C. (1994). Elicited imitation as a measure of second-language competence. In E. Tarone, S. M. Gass, & A. D. Cohen (Eds.), *Research methodology in second-language acquisition* (pp. 245–261). Mahwah, NJ: Erlbaum.

Bradlow, A. R., Akahane-Yamada, R., Pisoni, D. B., & Tohkura, Y. (1999). Training Japanese listeners to identify English /r/ and /l/: Long-term retention of learning in perception and production. *Perception and Psychophysics*, *61*, 977–985.

Bradlow, A. R., Pisoni, D. B., Akahane-Yamada, R., & Tohkura, Y. (1997). Training Japanese listeners to identify English /r/ and /l/: IV—Some effects of perceptual learning on speech production. *Journal of the Acoustical Society of America*, *101*, 2299–2310.

Burger, S., & Chrétien, M. (2001). The development of oral production in content-based second language courses at the University of Ottawa. *Canadian Modern Language Review*, *58*, 84–102.

Burger, S., Wesche, M., & Migneron, M. (1997). Late, late immersion: Discipline-based second language teaching at the University of Ottawa. In R. K. Johnson & M. Swain (Eds.), *Immersion education: International perspectives* (pp. 65–84). New York: Cambridge University Press.

Chastain, K. (1969). Prediction of success in audiolingual and cognitive classes. *Language Learning*, *19*, 27–39.
Daneman, M., & Newson, M. (1992). Assessing the importance of subvocalization during normal silent reading. *Reading and Writing: An Interdisciplinary Journal*, *4*, 55–77.
De Jong, N. (2005). Can second language grammar be learned through listening? An experimental study. *Studies in Second Language Acquisition*, *27*, 205–234.
DeKeyser, R. (2007). Skill acquisition theory. In B. VanPatten & J. Williams (Eds.), *Theories in second language acquisition* (pp. 97–113). Mahwah, NJ: Erlbaum.
Derwing, T. M., Munro, M. J., & Wiebe, G. (1998). Evidence in favor of a broad framework for pronunciation instruction. *Language Learning*, *48*, 393–410.
Di Biase, B. (1994). Innovative programs for learning Italian. *The Digest of Australian Languages and Literacy Issues*, *9*, 1–2.
Dulay, H. C., & Burt, M. K. (1973). Should we teach children syntax? *Language Learning*, *23*, 245–258.
Dulay, H. C., Burt, M. K., & Krashen, S. D. (1982). *Language two*. Oxford: Oxford University Press.
Ellis, N. C. (2007). The associative-cognitive CREED. In B. VanPatten & J. Williams (Eds.), *Theories in second language acquisition* (pp. 77–95). Mahwah, NJ: Erlbaum.
Erlam, R. (2006). Elicited imitation as a measure of L2 implicit knowledge: An empirical validation study. *Applied Linguistics*, *27*, 464–491.
Ervin-Tripp, S. (1974). Is second language learning like the first? *TESOL Quarterly*, *8*, 111–127.
Flege, J. E. (1995). Second-language speech learning: Theory, findings, and problems. In W. Strange (Ed.), *Speech perception and linguistic experience: Issues in cross-language research* (pp. 229–273). Timonium, MD: York Press.
Flege, J. E., MacKay, I. R. A., & Meador, D. (1999). Native Italian speakers' perception and production of English vowels. *Journal of the Acoustical Society of America*, *106*, 2978–2987.
Forsyth, A. (1990). Projet expérimental en anglais langue seconde à l'élémentaire au Nouveau-Brunswick [Experimental program in English as a second language in New Brunswick elementary schools]. *Education Canada*, *30*, 23–29.
Fraser, C., Bellugi, U., & Brown, R. (1963). Control of grammar in imitation, comprehension and production. *Journal of Verbal Learning and Verbal Behavior*, *2*, 121–135.
Gallimore, R., & Tharp, R. (1981). The interpretation of elicited imitation in a standardized context. *Language Learning*, *31*, 369–392.
Gary, J. O. (1975). Delayed oral practice in initial stages of second language learning. In M. K. Burt & H. C. Dulay (Eds.), *New directions in second language teaching, learning and bilingual education* (pp. 89–95). Washington, DC: TESOL.
Gathercole, S. E., Willis, C. S., Baddeley, A. D., & Emslie, H. (1994). The children's test of nonword repetition: A test of phonological working memory. *Memory*, *2*, 103–127.
Genesee, F. (1987). *Learning through two languages: Studies of immersion and bilingual education*. Rowley, MA: Newbury House.
Gibbons, J. (1986). The silent period: An examination. *Language Learning*, *35*, 255–267.
Goldschneider, J. M., & DeKeyser, R. (2001). Explaining the "natural order of L2 morpheme acquisition" in English: A meta-analysis of multiple determinants. *Language Learning*, *51*, 1–50.
Hauptman, P., Wesche, M., & Ready, D. (1988). Second language acquisition through subject-matter learning: A follow-up study at the University of Ottawa. *Language Learning*, *38*, 433–475.
Izumi, Y., & Izumi, S. (2004). Investigating the effects of oral output on the learning of relative clauses in English: Issues in the psycholinguistic requirements for effective output tasks. *Canadian Modern Language Review*, *60*, 587–609.
Janda, R. D., & Auger, J. (1992). Quantitative evidence, qualitative hypercorrection, socio-linguistic variables—And French speakers' 'eadhaches with English h/Ø. *Language and Communication*, *12*, 195–236.
Johnston, D. (1999). *Cool Edit 2000* [Computer software]. Phoenix, AZ: Syntrillium Software.
Krashen, S. D. (1985). *The input hypothesis: Issues and implications*. London: Longman.

Krashen, S. D. (1993). *The power of reading: Insights from the research*. Englewood, CO: Libraries Unlimited.
Krashen, S. D. (2003). *Explorations in language acquisition and use*. Portsmouth, NH: Heinemann.
Krashen, S. D., & Terrell, T. (1983). *The natural approach: Language acquisition in the classroom*. New York: Pergamon.
Lambacher, S. G., Martens, W. L., Kakehi, K., Marasinghe, C. A., & Molholt, G. (2005). The effects of identification training on the identification and production of American English vowels by native speakers of Japanese. *Applied Psycholinguistics*, *26*, 227–247.
Lightbown, P. M. (1985). Input and acquisition for second language learners in and out of classrooms. *Applied Linguistics*, *6*, 263–273.
Lightbown, P. M. (1992a). Can they do it themselves? A comprehension-based ESL course for young children. In R. Courchêne, J. St John, C. Thérien, & J. I. Glidden (Eds.), *Comprehension-based second language teaching* (pp. 353–370). Ottawa, Canada: University of Ottawa Press.
Lightbown, P. M. (1992b). Getting quality input in the second/foreign language classroom. In C. Kramsch & S. McConnell-Ginet (Eds.), *Text and context: Cross-disciplinary and cross-cultural perspectives on language study* (pp. 187–197). Lexington, MA: D. C. Heath.
Lightbown, P. M., Halter, R. H., White, J., & Horst, M. (2002). Comprehension-based learning: The limits of "do it yourself." *Canadian Modern Language Review*, *58*, 427–464.
Long, M. H. (1980). Inside the "black box": Methodological issues in classroom research on language learning. *Language Learning*, *30*, 1–42.
Long, M. H. (1981). Input, interaction, and second-language acquisition. *Annals of the New York Academy of Sciences*, *379*, 259–278.
Long, M. H. (1991). Focus on form: A design feature in language teaching methodology. In K. de Bot, D. Coste, R. Ginsberg, & C. Kramsch (Eds.), *Foreign language research in cross-cultural perspectives* (pp. 39–52). Amsterdam: Benjamins.
Macnamara, J. (1973). Nurseries, streets, and classrooms: Some comparisons and deductions. *Modern Language Journal*, *57*, 250–255.
McCandless, P., & Winitz, H. (1986). Test of pronunciation following one year of comprehension instruction in college German. *Modern Language Journal*, *70*, 355–362.
McDade, H., Simpson, M., & Lamb, D. (1982). The use of elicited imitation as a measure of expressive grammar: A question of validity. *Journal of Speech and Hearing Disorders*, *47*, 19–24.
Munro, M. J., & Derwing, T. M. (1999). Foreign accent, comprehensibility and intelligibility in the speech of second language learners. *Language Learning*, *49*, 285–310.
Musumeci, D. (1997). *Breaking tradition: An exploration of the historical relationship between theory and practice in second language teaching*. New York: McGraw Hill.
Neufeld, G. G. (1978). On the acquisition of prosodic and articulatory features in adult language learning. *Canadian Modern Language Review*, *34*, 161–174.
Palmer, H. E. (1968). *The scientific study and teaching of languages*. Oxford: Oxford University Press. (Original work published 1917)
Paradis, C., & LaCharité, D. (2001). Guttural deletion in loanwords. *Phonology*, *18*, 255–300.
Pierrehumbert, J. B. (2003). Probabilistic phonology: Discrimination and robustness. In R. Bod, J. Hay, & S. Jannedy (Eds.), *Probabilistic linguistics* (pp. 177–228). Cambridge, MA: MIT Press.
Postovsky, V. (1974). Effects of delay in oral practice at the beginning of second language learning. *Modern Language Journal*, *58*, 229–239.
Rayner, K. (1988). Word recognition cues in children: The relative use of graphemic cues, orthographic cues, and grapheme-phoneme correspondence rules. *Journal of Educational Psychology*, *80*, 473–479.
Ready, D., & Wesche, M. (1992). An evaluation of the University of Ottawa's sheltered program: Language teaching strategies at work. In R. Courchêne, J. St John, C. Thérien, & J. I. Glidden (Eds.), *Comprehension-based second language teaching* (pp. 389–405). Ottawa, Canada: University of Ottawa Press.
Robinson, P., & Ellis, N. C. (Eds.). (2008). *Handbook of cognitive linguistics and second language acquisition*. London: Routledge.

Rvachew, S. (1994). Speech perception training can facilitate sound production learning. *Journal of Speech and Hearing Research*, *37*, 347–357.
Schmidt, R. (1990). The role of consciousness in second language learning. *Applied Linguistics*, *11*, 129–158.
Sheldon, A., & Strange, W. (1982). The acquisition of /r/ and /l/ by Japanese learners of English: Evidence that speech production can precede speech perception. *Applied Psycholinguistics*, *3*, 243–261.
Slobin, D. I., & Welsh, C. A. (1973). Elicited imitation as a research tool in developmental psycholinguistics. In C. A. Ferguson & D. I. Slobin (Eds.), *Studies of child language development* (pp. 485–497). New York: Holt, Rinehart, & Winston.
Strevens, P. (1960). Spectra of fricative noise in human speech. *Language and Speech*, *3*, 32–49.
Swain, M. (1985). Communicative competence: Some roles of comprehensible input and comprehensible output in its development. In S. M. Gass & C. Madden (Eds.), *Input in second language acquisition* (pp. 235–253). Rowley, MA: Newbury House.
Trofimovich, P. (2008). What do second language listeners know about spoken words? Effects of experience and attention in spoken word processing. *Journal of Psycholinguistic Research*, *37*, 309–329.
Trofimovich, P., & Baker, W. (2006). Learning second-language suprasegmentals: Effect of L2 experience on prosody and fluency characteristics of L2 speech. *Studies in Second Language Acquisition*, *28*, 1–30.
VanPatten, B. (2004). *Processing instruction: Theory, research, and commentary*. Mahwah, NJ: Erlbaum.
White, L. (1987). Against comprehensible input: The input hypothesis and the development of second-language competence. *Applied Linguistics*, *8*, 95–110.
Winitz, H., Gillespie, B., & Starcev, J. (1995). The development of English speech patterns of a 7-year-old Polish-speaking child. *Journal of Psycholinguistic Research*, *24*, 117–143.
Winitz, H., & Reeds, J. A. (1973). Rapid acquisition of a foreign language (German) by the avoidance of speaking. *International Review of Applied Linguistics*, *11*, 295–316.
Wode, H. (1980). *Learning a second language: An integrated view of language acquisition*. Tübingen: Gunter Narr Verlag.
Wode, H. (1996). Speech perception and L2 phonological acquisition. In P. Jordens & J. Lalleman (Eds.), *Investigating second language acquisition* (pp. 321–353). Berlin: Mouton de Gruyter.

APPENDIX

STIMULUS SENTENCES USED IN THE ELICITED IMITATION TASK

1. Sam *is* a clown.
2. *His hair is* blue and *his* nose *is* red.
3. Sam *has* a dog.
4. *His* **dog's** name *is* Tom.
5. Where *is* **Sam's** *hat*?
6. *Oh* no! Tom *is eating* **Sam's** *hat.*

Note. Word tokens used to calculate English /h/ production accuracy are in italics, whereas word tokens used to calculate English possessive /z/ production accuracy are boldfaced.