DOI: 10.56794/VSSR.6(212).70-84

Second Language Speech Production as Seen from Levelt's Model Adaptation

Đỗ Thị Huyền Thanh[*], Đặng Thị Minh Tâm[*]

Received on 30 September 2022. Accepted on 10 November 2022.

Abstract: Second language (L2) speech production plays an important role in language teaching and learning, as well as in machine learning. Among several promising paradigms, Levelt's model is still considered suitable for further studies. This paper reviews Levelt's speech production model and its adaptation to L2 speech production. Firstly, Levelt's modular model of first language (L1) speech production is revisited as a basis for understanding its adapted versions for L2 speech production. Next, details of stages of L2 speech production (Conceptualisation, Formulation, Articulation, and Monitoring) are presented in comparison with L1 speech production. Then, recent research related to L2 speech production is reviewed. The document analysis method is employed to find out the differences between L1 and L2 speech production, based on which implications for L2 speech production and acquisition are discussed. It is hoped that understanding cognitive processes involved in producing an L2 speech can facilitate teachers in teaching L2 speaking skills thanks to knowing how to choose suitable teaching materials and methods and develop valid instruments to measure learners' oral competence.

Keywords: Speech production, conceptualisation, formulation, articulation, monitoring.

Subject classification: Linguistics

1. Introduction

There is no doubt that one of the most important goals of learning a second language is being able to speak the language fluently. Understanding how speech is produced is the key to improving the teaching of speaking skills and helping learners improve their oral language ability. To fully understand how L2 speech production mechanism works, we first need to understand how L1 speech is produced.

Most theories of L1 production follow two main trends: the spreading activation theory (e.g., Dell, 1986) and the modular theory of speech processing (e.g., Levelt, 1989, 1993). The first difference between these theories is that in spreading activation theories, backward

[*] Hanoi University of Industry.
Email: thanhdth@haui.edu.vn

activation from a subordinate level to the superordinate level is possible, whereas modular models only allow activation to spread one-way forward. Another difference relates to syntactic and phonological encoding. Spreading activation theories postulate the so-called "frame-slot models of production" indicating that frames (with slots to be filled) for sentences or phonetic representations are constructed first, and then suitable words or phonetic features will be selected and inserted into the slots accordingly. In contrast, modular models are lexically driven, arguing that words activate syntactic encoding, followed by phonological encoding. However, Dell and fellows seemed to accept some of the most crucial arguments of Levelt. "In a later article, Dell, Juliano, and Govindje (1993) gave up the claim that activation can spread backward from the phonological to the lexical level, and they concluded that there is no need for the frame-slot mechanism and generative rules in syntactic and phonological encoding" (Kormos, 2006, p. 6). This is not to imply that the model of Level is the supreme and unique, rather we assume that it bears some promising features that can expel the learning and teaching of L2 in Vietnam in the context of global integration. Therefore, this paper will focus only on Levelt's modular model of L1 speech production as a basis for understanding the processes underlying L2 speech production. Four stages of L2 speech production, namely the Conceptualisation, Formulation, Articulation, and Monitoring, are discussed in detail and compared with those in L1 speech production. Then, recent research related to L2 speech production is reviewed using the document analysis method to find out the differences between L1 and L2 speech production, based on which implications for L2 speech production and acquisition are discussed.

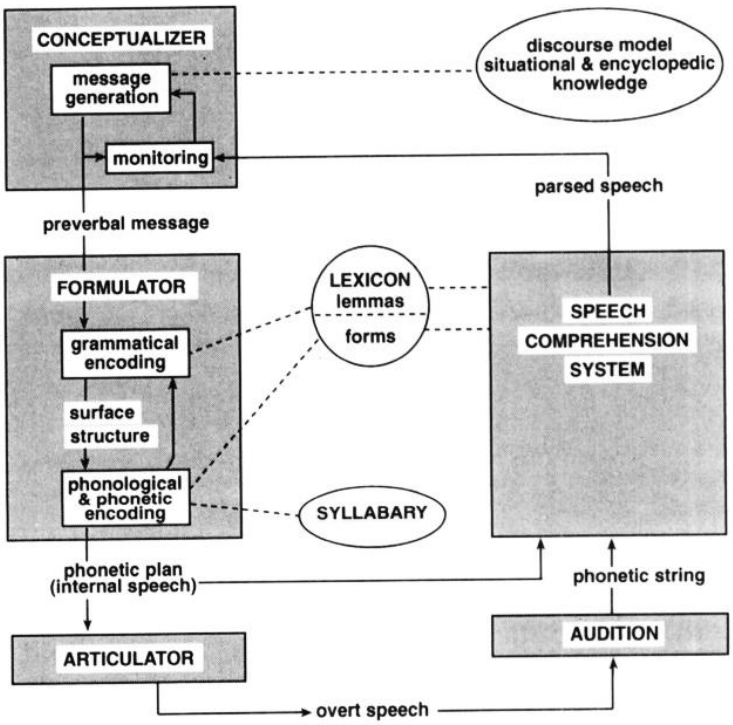
2. L1 speech production

A highly influential and well-established model of speech production is that of Levelt (1989, 1995, 1999), which is labelled "*the blueprint for the native speaker*" (Levelt, 1989, p.9). Levelt's model posits that speech production is modular and includes four independent and sequential stages, namely *Conceptualisation*, *Formulation, Articulation*, and *Monitoring* with various sub-processes involved in each.

According to Levelt, there are various processing components involved in the production of speech, including processing modules and knowledge stores. (see Figure 1). The primary modules within this model are the Conceptualiser, the Formulator, the Articulator, and the Speech Comprehension System. According to Levelt (1989, 1995, 1999), the Conceptualiser has two functions, namely generating preverbal messages, and monitoring the whole of speech production. Next, the Formulator is responsible for formulating the language representation of the message (i.e., encoding it grammatically, and phonologically) to produce a phonetic plan (internal speech). The Articulator is then in charge of executing the phonetic plan and converting it into overt speech. Finally, the Speech Comprehension System allows the monitoring of speech production to take place by making both internal and overt speech available to the conceptual system. Each of these components receives some kind of input and produces a certain kind of output, which then serves as input for the

next component. Furthermore, the information flows unidirectionally between the components (e.g., forward from the Conceptualiser to the Formulator, but not backward from the Formulator to the Conceptualiser). Briefly, producing speech involves the speaker's conceptualising the message, encoding the message, and finally articulating it. Monitoring can occur during each of these stages. These production stages are accomplished by the speaker accessing various knowledge sources. According to Levelt's (1995) model, there exist three knowledge stores: *the store of World Knowledge*, *the Lexicon*, and *the Syllabary*. The first store contains the speaker's external and internal knowledge of the world (also called encyclopedic knowledge), which is accessed for conceptualisation. The second store, *the Lexicon*, is composed of two parts, the lemma and the lexeme. In the lemma, a lexical entry's meaning and syntax are essential for grammatical encoding and are represented as declarative knowledge, whereas the lexeme consists of procedural knowledge of a lexical entry's morphology and phonology, which are used for phonological encoding. Finally, *the Syllabary* contains gestural scores that are used to produce the syllables of the actual speech.

Figure 1: Processing Components Involved in Generation of Speech



Source: W. J. M. Levelt, 1995, p.14.

In the *Conceptualisation* stage, a preverbal message is generated when speakers select information from their world knowledge and use this to convey in their message. The

generation of preverbal messages from intention includes two substages: macro-planning and micro-planning. Macro-planning involves breaking a communicative intention down to one or more individual speech acts and ordering them for expression. Following the macro-planning stage of conceptualisation, the micro-planning stage then involves further shaping the speech acts into the format of a preverbal message. According to Levelt (1989), macro-planning and micro-planning are two incremental stages which can alternate with one another or occur simultaneously during the conceptualisation phase. It should also be noted that the preverbal message contains conceptual information (e.g., semantics, style, register) that is not yet linguistic, and will constitute the input for the next processing component in Levelt's model, *the Formulator*, to work on.

In *the Formulation* stage, the preverbal message (a conceptual structure) is converted into a phonetic plan (a linguistic structure) through two processes: grammatical and phonological encoding. *Grammatical encoding* involves accessing a lexical item's lemma information (i.e., meaning and syntax) stored in the lexicon, and relevant syntactic building procedures to produce a surface structure (see Figrure 1). *Phonological encoding* involves retrieving lexeme information (i.e., morphology and phonology) and syntactic activation of lexical forms. The output of phonological encoding is a phonetic plan, which is also called internal speech because it is not yet overt speech but "an internal representation of how the planned utterance should be articulated - a programme for articulation" (Levelt, 1989, p.12).

In the third stage of speech production, Articulation, the phonetic plan is realised as the sounds and syllables of speech are produced. The outcome of the articulation stage is then the overt speech that a speaker produces.

Finally, Monitoring, which is done by the Conceptualiser, involves checking the accuracy and appropriateness of the output of the three modules in the model. While the other three stages (conceptualisation, formulation, and articulation) operate independently in a sequential input-output manner (i.e., the output of one stage functions as input for the next stage), monitoring can occur at any point within these stages. When the monitoring system inside the Conceptualiser detects any meaning or form deviations between the original intention and the parsed speech, it may interrupt the speech stream, reformulate the preverbal message, and send a repair message to the Formulator. This is one source of a speaker's false starts, hesitations, and self-repairs.

In brief, the whole process of L1 speech production in Levelt's (1989, 1995, 1999) model can thus be summarised as going through three main stages: generating a preverbal message (*Conceptualisation*), creating a phonetic plan through grammatical and phonological encoding (*Formulation*), and producing overt speech (*Articulation*) with *Monitoring* taking place at any of these stages. These stages of speech production involve the work of three modules: the Conceptualiser, the Formulator, and the Articulator. Each module's operation is independent of the other modules with its own characteristic input, and there is no direct exchange of information between them. Feedback between modules, however, is provided through monitoring. It is important to note that different stages of speech production can take place simultaneously, provided that the processing in one of the modules is sufficiently

automatised for the other to proceed uninhibited. For example, when *the Conceptualiser* has passed its output to the next component (*the Formulator*), it can start with another piece of input at the same time as the previous output is being simultaneously encoded by *the Formulator*. However, this can only occur if the activities of *the Formulator* are sufficiently automated to allow the speaker's attention to move forward with conceptualisation, rather than being diverted into lexical retrieval and grammatical encoding. For L1 speakers, the activities of the Formulator are largely automatised. While message conceptualisation and monitoring require the L1 speaker's attention and memory resources, grammatical and phonological encoding and articulation are usually automatic processes that require little attention and can take place in parallel (Levelt, 1989; Poulisse, 1997). These features of parallel processing and automaticity together allow for speedy real-time L1 language production.

3. L2 speech production

Most researchers in the field of second language acquisition (SLA) are interested in understanding the processes involved in L2 speech production and how these processes are different from L1. Among those, De Bot (1992), Kormos (2006), and Segalowitz (2010) are seen to have best adapted Levelt's (1989, 1995, 1999) modular model of L1 speech production to the conditions governing L2 speech production.

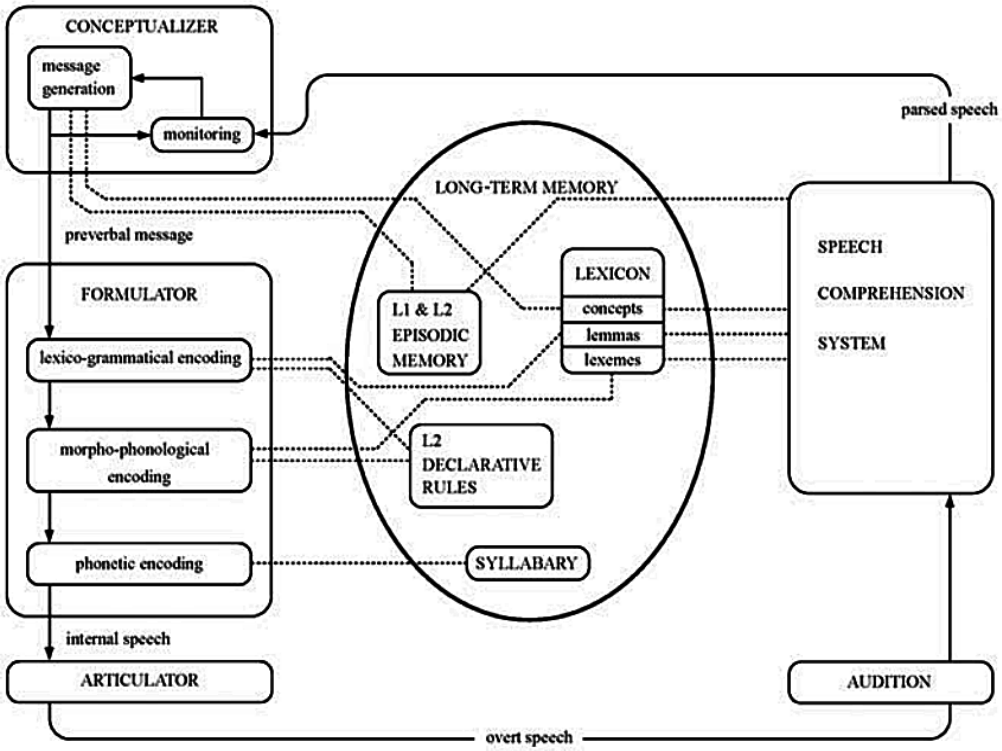
Firstly, De Bot (1992) adapted Levelt's model to account for bilingual speech production by introducing additional requirements beyond those of a monolingual processing model. Regarding *the Conceptualiser*, De Bot (1992, p.21) argued that the first of the two production phases of conceptualisation (macro-planning) is not language-specific, whereas the second phase (micro-planning) is language-specific. In other words, macro-planning may involve activating normally identical relevant concepts or shared between languages, whereas micro-planning is likely to take place in the intended language only. For De Bot, information about the language in which an utterance is produced is, therefore, specified in the preverbal message which is the outcome of the Conceptualiser and the input to the Formulator.

In terms of the Formulator module, De Bot (1992) argued that the way L2 formulator operates in L2 production is almost the same as the L1 formulator does in L1 production. In both cases, the preverbal plan is converted into a phonetic plan. However, De Bot also indicated that each language possesses its own distinctive micro-planning and formulator, but these formulators draw on a single lexical store (the lexicon) where both L1 and L2 lexical items are stored together. De Bot thus proposed the idea of parallel phonetic plans in the L1 and the L2. The respective formulators send their own phonetic plan to *the Articulator*. For De Bot, the Articulator is not language-specific and has "an extensive set of sounds and pitch patterns from both languages" (De Bot, 1992, p.17) that are used to produce overt speech.

Later on, Kormos (2006) provided a comprehensive model of L2 speech production (see Figure 2) in relationship to Levelt's (1989, 1999) model of L1 speech production. Like

De Bot (1992), Kormos (2006) acknowledged that production mechanisms in both the L1 and the L2 are similar, consisting of separate specialist processing modules (i.e., *the Conceptualiser*, *the Formulator*, and *the Articulator*). Furthermore, she acknowledges that L2 speech production in two modules can take place simultaneously, provided that production in one of the modules is sufficiently automatic to allow the production in the other to proceed unimpeded.

Figure 2: Model of Bilingual Speech Production



Source: J. Kormos, 2006, p.168.

However, Kormos (2006) disagreed that different types of knowledge are stored in separate knowledge stores and proposes some modifications in knowledge stores to account for L2 speech production. In particular, L2 speech production draws on one large memory store, namely the Long-Term Memory, instead of three knowledge stores (*the store of World Knowledge, the Lexicon, and the Syllabary*) as presented in Levelt's L1 speech production model (see Figure 1). This long-term memory store is composed of several sub-stores: an episodic memory, a semantic memory (the lexicon), a syllabary, and a declarative knowledge memory. The episodic memory contains temporally organised experiences in one's life; the semantic memory consists of concepts, lemmas (syntactic information), and

lexemes (morpho-phonological information); and the syllabary contains the automatised gestural scores which are used to produce syllables for internal speech. It is noted that all these three knowledge stores are shared between the L1 and the L2. Finally, Kormos (2006) argued for the existence of a fourth and important knowledge store in L2 production, a declarative knowledge memory, which contains information about syntactic and phonological rules in L2. This memory of declarative knowledge is L2-specific because, as Kormos points out, for L2 speakers, many syntactic and phonological rules in L2 are not yet automatised and are stored as declarative knowledge, while in L1 production these rules are almost automatic. Kormos also notes that, except for the addition of the new declarative knowledge store and the reorganisation of the four knowledge sub-stores into one large memory store (*the Long-Term Memory*), there is no significant difference between the bilingual speech production model and the one constructed for monolingual speakers.

Following De Bot (1992) and Kormos (2006), Segalowitz (2010) proposes an updated and integrated model of the L2 speaker by adapting Levelt's (1999) L1 speech production model and incorporating De Bot's (1992) amendments regarding L2 speech production. Specifically, in this model, Segalowitz identifies seven potentially critical points where underlying processing issues could lead to dysfluencies in L2 speech. These points are called "fluency vulnerability points" and are marked with f symbols. The seven fluency vulnerability points include: (f1) microplanning, (f2) grammatical encoding, (f3) lemma retrieval, (f4) morpho-phonological encoding, (f5) phonetic encoding, (f6) articulation, and (f7) self-perception (for details, see Segalowitz, 2010, p.9). Encountering difficulties at any of these points may lead L2 speakers to interrupt their speech fluidity.

In brief, researchers agree that L2 production is similar to L1 production in that it also has four important processing components: conceptualisation, formulation, articulation, and monitoring. However, it is also agreed that L2 production is distinct from L1 production in certain aspects due to the influence of the L1 on the L2, the incomplete nature of L2 knowledge, learners' access to L2 resources, as well as the limited attentional capacity that speakers bring to the task of L2 production (De Bot, 1992; Kormos, 2006). Differences between L1 and L2 speech production are also due to the vulnerability points in L2 speech processing (Segalowitz, 2010) which L1 speakers do not encounter because they can resort to automatic and parallel processing. These differences explain why L2 speakers need to resort to the use of serial processing, L1 transfer and the use of communication strategies during real-time L2 speech production. Details of each stage of L2 speech production will be discussed below.

3.1. L2 conceptualisation

Conceptualisation in an L2, like in L1, involves the planning of what a speaker wants to say to realise a communicative intention. In this stage, the speaker makes decisions about the content of the preverbal message and its organisation by selecting information from his or her world knowledge and organising it into an initial pre-verbal structure before

deciding on the language that will be used to express it. In line with Levelt's model (1989, 1995, 1999), conceptualisation in an L2 can be broken down into two substages, namely macro-planning and micro-planning (De Bot, 1992; Segalowitz, 2010) with the preverbal message as the output. The question that Kormos (2006) raises about this stage is "whether speakers formulate parallel speech plans - a plan for L1 and another one for L2 - or a single speech plan in which each concept is labelled with a language tag" (p.xx). The idea of parallel preverbal plans in L1 and L2 was proposed by De Bot (1992) but was rejected a year later by De Bot and Schreuder (1993). Since then, most researchers argue for a single preverbal plan that specifies both conceptual information and the language to be used to express the message. In brief, except the language tag specified in the preverbal message, conceptualisation in the L2 remains the same as in L1 speech production.

3.2. L2 formulation

In the Formulation stage, the preverbal message activates items in the mental lexicon corresponding to different chunks of the intended message, and this preverbal plan is then formulated into a phonetic plan (internal speech) through lexico-grammatical, morpho-phonological, and phonetic encoding.

In lexico-grammatical encoding, the conceptual specifications in the preverbal plan activate corresponding lemmas in the lexicon (i.e., lexical encoding), then this lexical-syntactic information is used to build up the surface structure (i.e., syntactic encoding). Regarding lexical encoding, Kormos (2006) support the position that (1) not only L2 but L1 lemmas are also activated to some extent, and (2) both L1 and L2 activated lemmas compete for selection, but the lemmas whose features best match the conceptual specifications and the language cue will be selected. Regarding syntactic encoding, Kormos (2006, p.171), like De Bot (1992), argues that L2 production is not significantly different from L1 production in that it is "lexically driven" and comprises several sequential sub-phases. As in L1 production, L2 syntactic encoding involves first the activation of syntactic features associated with a lexical item; then employing syntactic encoding procedures to build up phrases and clauses, and arranging these phrases into an appropriate sequence for an utterance. At this stage, like L1 speakers, proficient L2 speakers normally have automatic access to procedural knowledge of syntactic and morphological rules. However, for lower proficiency level learners, when a form is not fully procedualised, learners may resort to the use of communication and transfer strategies which require additional attention to the formulation of speech and prevent the learners from parallel processing in other modules. As a result of searching for language to express their ideas, L2 speakers need to serially process their speech one stage at a time from conceptualisation to formulation, causing a breakdown in fluency and a decrease in speech rate.

In the next phase, the morpho-phonological form of the lexical items is activated and syllabified in its syntactic context, and the parameters for loudness, pitch, and duration are set. Kormos (2006) argues that the phonological form of words in the non-selected language is also

activated, and both L1 and L2 lexemes compete for selection. She proposes that the processing mechanisms of L2 phonological encoding work similarly as in L1 speech production, but that key adjustments need to be made to account for the fact that the L2 speaker already speaks an L1 and that aspects of the L1 will impact the processing of their L2.

In the phonetic encoding phase, the phonemes of words are activated in a serial manner. According to Kormos (2006), representations of these L1 and L2 phonemes are stored together in a single network in the syllabary. Kormos suggests that L2 learners may often use L1 phonemes in place of similar L2 phonemes, or they may apply phonological processes associate with the pronunciation of the L1 in encoding to L2 phonology. This results in the accent in which they speak the L2.

In the formulation stage of speech production, it becomes obvious that the difference between monolingual and bilingual speech processing involves the influence of L1 on the L2. This difference seems to be unavoidable as most of the knowledge stores are shared in L1 and L2, and thus both L1 and L2 linguistic items compete for selection. This competition can result in unintentional switching between L2 to L1 forms. Another result of the competition between L1 and L2 forms is the transfer of L1 production rules when a production process that is appropriate to the L1, but not to the L2, is employed.

3.3. L2 articulation

Articulation is the third processing module in both L1 and L2 speech production. At the articulation stage, articulatory gestures, or the physical motor skills involved in producing syllables, are retrieved and activated. This results in the overt speech of L2 learners. In L1 speech production, syllables are assumed to be the basic units of articulatory execution, and the phonetic plan is composed of numerous syllable programmes. For bilingual speakers, however, whether these syllable programmes are automatised or not is likely to depend on the speaker's L1 as well as their level of proficiency in their L2. In line with De Bot's (1992) view, Kormos (2006) hypothesises that lower proficiency L2 speakers are largely dependent on L1 syllable programmes, whereas higher proficiency L2 speakers can engage separate sets of motor skill programmes for their L1 and their L2.

3.4. L2 monitoring

The fourth and final process in L2 speech production is monitoring, which deals with learners' attention to their output to check the accuracy and appropriateness of the output from each of the three primary speech production modules (conceptualisation, formulation, and articulation). According to Kormos (2006), monitoring in L2 speech production involves the same mechanisms as speech comprehension and proceeds similarly to that in L1. Specifically, there are three monitor loops that are responsible for checking the outcome of production processes. The first loop evaluates the output conceptually with the original intentions of the preverbal plan. The second loop attends to the accurate formulation of the

message in linguistic form by engaging internal speech and focusing attention on the phonetic plan before it was articulated. The last loop involves attention to the overt speech to check for pronunciation or other articulation problems. Upon detecting any trouble in the output regarding any of these loops, "the monitor issues an alarm signal, which, in turn, triggers the production mechanism for a second time starting from the phase of conceptualisation" (Kormos, 2006, p.173). The speaker then stops the speech stream, and either modifies messages, repairs the utterance or reformulates it entirely, depending on the nature of the issues, the speakers' L2 resources, and the time constraints imposed by the context. In all of these cases, breakdowns in speech processing may occur, resulting in pauses and repetitions to buy themselves time to meet these additional processing demands.

Despite similar processing mechanisms, monitoring in the L1 differs from monitoring in the L2 in several ways, mostly due to attentional demands (Kormos, 2006). While formulation and articulation in L1 production are largely automatic (attention-free) and can operate in parallel with conceptualisation and monitoring, L2 speech processing requires attention and serial processing at almost all levels. With limited attentional resources, L2 speakers, therefore, have less attention to spare for monitoring than L1 speakers, and they must decide what to prioritise (Kormos, 2011). As all aspects of speech production require attention at some point, instruction must be designed to support learners by ensuring that attention is alternated between relevant aspects of their developing L2 speech processing capacity (Skehan, 2009).

3.5. Factors impacting L2 speech production

In the field of SLA, it is generally agreed that there is a golden period (between the ages of 11 to 18) in which learners can acquire a language implicitly (Lightbown & Spada, 2013). After this period ends, language is acquired in the same way as other skills through proceduralisation of declarative knowledge (which enables the automatisation of encoding processes), and memorisation of formulaic expressions. As discussed above, the efficiency of L2 speech production can be affected by L1 influence because both L1 and L2 lexical items are activated and competed for selection. Moreover, producing an utterance involves fast and efficient coordination of many cognitive processes such as conceptual planning, lexical and grammar encoding, articulation, and monitoring. It is assumed that L2 speech production is characterised by utterance fluency and depends on the extent to which these cognitive processes are automatic (Hilton, 2008; Kormos, 2006; Segalowitz, 2010). For low-proficiency L2 speakers with limited attention and incomplete linguistic knowledge of the L2, the encoding processes underlying L2 speech production require serial processing instead of automatic processing as in the L1, causing L2 speech to be less fluent with more pausing and dysfluencies (Kormos, 2006). Research has shown that factors impacting L2 speech production the most include vocabulary knowledge (vocabulary size and lexical retrieval speed) (De Jong et al., 2013; Koizumi et al., 2013; Liu, 2020; Uchihara & Saito, 2019), and the degree of proceduralisation of L2 linguistic knowledge (Segalowitz, 2010).

Regarding the relationship between L2 vocabulary knowledge and speech production, studies have found significant correlation between learners' vocabulary knowledge and their speaking proficiency (De Jong et al., 2013; Hilton, 2008; Koizumi et al., 2013; Uchihara & Saito, 2019). De Jong et al. (2013) found that most objective measures of fluency are affected by both linguistic knowledge (including vocabulary) and linguistic processing skills (e.g., lexical retrieval speed). Hilton (2008) also confirmed the important role of lexical competence in spoken L2 fluency when pointing out that the lack of lexical knowledge, or limited access to vocabulary, would lead to of the most serious dysfluencies. Koizumi et al. (2013) found that 60% of the variance in speaking proficiency can be explained by vocabulary size. Uchihara and Saito (2019) investigated the relationship between productive vocabulary knowledge and L2 speaking performance and demonstrated that L2 learners' fluency can be predicted by their productive vocabulary knowledge. The study results suggested that that more developed lexical knowledge resulted in easier retrieval of L2 words, which then led to more fluent speech. More recently, Liu (2020) examined the relationship between L2 learners' lexical access (measured by vocabulary size and lexical retrieval speed), and three dimensions of their speaking performance (fluency, accuracy, and complexity). The results showed that lexical access was highly correlated to all three dimensions of L2 speech and that vocabulary size and lexical retrieval speed could be important predictors of fluency, accuracy, and complexity. The results of these previous studies confirm the role of lexical access in second language speech production and raise the importance of helping learners to build up a large L2 lexicon, getting it ready for optimal access during the online encoding process of L2 speech production.

Regarding the proceduralisation issue of L2 linguistic knowledge, a great deal of studies has examined how to improve the automaticity of linguistic knowledge and skills through practice. As mentioned above, L2 speech production (characterised by L2 fluency) is subjected to many factors, including the degree to which the cognitive processes are automatic (Hilton, 2008; Kormos, 2006; Segalowitz, 2010). The difficulties associated with linguistic encoding during real-time communication naturally lead to a slower and more disrupted L2 speech with a range of dysfluencies like silent pauses, fillers (e.g., *uh* and *um*), repetitions, false starts, and self-repairs (De Jong et al., 2013; Derwing et al., 2009). Therefore, to improve L2 learners' speech fluency, it is important to facilitate the conversion of declarative knowledge into procedural knowledge as a key to enable parallel processing of different cognitive processes underlying L2 speech production because when knowledge becomes proceduralised, it is accessed automatically, effortlessly, and efficiently. It is suggested that practice has a role to play in this process, in that practice can lead to in changes in interlanguage by creating more chunks of information that are available for automatic processing.

Among different forms of practice, task repetition is one task implementation variable that has been investigated in numerous studies. According to Bygate and Samuda (2005), task repetition means 'repetitions of the same or slightly altered tasks - whether whole tasks, or parts of a task (p.43). Previous research has consistently shown that task repetition

is widely used in L2 classrooms and has beneficial effects for improving L2 performance in general and L2 fluency in particular (Ahmadian, 2011, 2012; Ahmadian, Masouri & Ghominejad, 2017; Bygate & Samuda, 2005; N. De Jong & Perfetti, 2011; Lambert, Aubrey & Leeming, 2020; Lambert et al., 2017). The impact of task repetition on L2 production has been mostly considered about three major stages of Levelt's model of speech production (i.e., conceptualisation, formulation, and articulation). *First,* having a prior performance means a lot of relevant work on conceptualisation, formulation, and articulation has been done before the speakers perform the task for the second time (Bygate & Samuda, 2005). Task repetition, thus, eases online processing demands of the task being performed. *Second*, repetition of tasks can provide learners with the opportunity to familiarise themselves with task demands, strengthen form-meaning connections, proceduralise relevant linguistic knowledge on the first performance, and make it available for automatic processing during subsequent performances. Repeated practice facilitates automatised access to linguistic knowledge and frees up more attentional resources so that the speakers can devote much of their cognitive resources to produce more fluent speech on subsequent performances.

Besides this, the impact of planning as another task implementation factor has been the topic of interest in numerous task-based studies (Ellis, 2009; Bui & Huang, 2016; Bui, 2014; Skehan, 2009; Wang, 2014). It is assumed that planning would influence L2 task performance because it allows some conceptualising work to be done before the task performance itself. According to Ellis (2009), there are three kinds of planning: rehearsal, pre-task planning, and within-task planning. Rehearsal provides learners opportunities to complete a task at least once before actual performance. Pre-task planning (strategic planning) allows learners to plan content or language before performing the actual task, whereas within-task planning (online planning) occurs when time is available during speech production. Bui (2014) argues that all three forms of planning prepare learners ready to do a task and research has shown effectiveness of those planning form on learners' performances in terms of complexity, accuracy, and fluency. Because of limited attention capacity, there are trade-offs among these aspects of language performance when L2 speakers focus on one aspect of performance. Recently, Lambert et al. (2020) investigated the relative impact of four task preparation options (same task repetition, parallel task repetition, L1 planning, and L2 planning) on L2 learners' speech production. The study revealed different effects of the four kinds of task preparation, and then suggested that different preparatory options be sequenced in a way that might support L2 learners' speech production by alternately easing the conceptualisation, formulation, and monitoring demands during task performance.

In brief, it can be seen that recent research investigating L2 speech production still based on Levelt's model of L1 speech production or its L2 adapted versions as a theoretical framework with reference to major stages of speech production. However, there remain unanswered questions about the model, like whether it can accommodate the changes in language development. De Bot, K., Schmid, M. S., & Lowie, W. (2011, p.2) postulated that "*things tend to become more problematic when these models are applied to non-static scenarios, such as linguistic change, language development, and multilingualism*". They also

suggested that "language development might be too complex and unpredictable to be captured by models based on linear and hierarchical assumptions, and that it might be time to consider a radical change of scientific paradigm" (De Bot et al., 2011, p.4). Researchers (e.g., De Bot, 2008) have recently turned to Complexity Theory or Dynamic Systems Theory to account for dynamic complexity of language development and doubted to what extent Levelt's model would hold. Another issue is whether the model is applicable with the development of artificial intelligence and digital computers as the primary metaphor for second language development.

4. Conclusion

The past two decades have witnessed the considerable development of bilingualism research, notably well-structured models like Levelt's modular model and its adapted versions in L2. These models are of great contribution to our knowledge of how L2 speech is made and how the underlying processes are different from those in L1. Despite contradictory opinions, Levelt's model remains one of the most influential and well-referenced theories in L2 speech research. This paper has reviewed theories of speech production, outlining the primary constructs to understand how speech is produced. It begins with Levelt's model of L1 production, then compared models of L1 and L2 speech production regarding processing mechanisms underlying each stage of conceptualisation, formulation, articulation, and monitoring. While it is agreed that most aspects of L2 production can be explained by the model of L1 production proposed by Levelt (1989, 1995, 1999), there are several differences between L1 and L2 speech production due to the influence of the L1 on the L2, the speaker's deficits in L2 linguistic knowledge, the degree of automaticity and the limited attentional capacity of L2 speakers. These issues explain why L2 speech contains more hesitations and is generally not as smooth and rapid as L1 speech. As such, improvement in speech fluency can be attributable to the efficient functioning of speech production mechanisms, including the automaticity of encoding processes. In other words, L2 fluency development can be subjected to the extent to which the cognitive processes underlying speech production are automatic (Hilton, 2008; Kormos, 2006; Segalowitz, 2010). L2 learners can speak the language fluently like L1 speakers do when the access to linguistic resources is largely automatised, which allows them to conceptualise a message and formulate it in linguistic forms at the same time (parallel processing). Therefore, to improve L2 learners' speech fluency, it is important to facilitate the conversion of declarative knowledge into procedural knowledge to enable parallel processing of different cognitive processes underlying L2 speech production. It is suggested that practice has a role to play in this process and teachers are to choose suitable teaching materials and methods, and to develop valid tools to assess learners' oral competence to make their teaching of the L2 speaking skill more efficiently. Future research can explore language development from a dynamic approach (e.g., Complexity Theory and Dynamic Systems Theory) or using computer modelling to gain different insights.

References

1. Ahmadian, M. J. (2011), "The Effect of 'Massed' Task Repetitions on Complexity, Accuracy, and Fluency: Does It Transfer to a New Task?", *The Language Learning Journal*, No. 39(3), pp.269-280.
2. Ahmadian, M. J. (2012), "Task Repetition in English Language Teaching", *English Language Teaching Journal*, No. 66(3), pp.380-382.
3. Ahmadian, M. J., Masouri, A., & Ghominejad, S. (2017), "Language Learners' and Teachers' Perceptions of Task Repetition", *English Language Teaching Journal*, No. 71(4), pp.467-477.
4. Bui, H.Y.G. (2014), "Task Readiness: Theoretical Framework and Empirical Evidence from Topic Familiarity, Strategic Planning, and Proficiency Levels", in P. Skehan (ed.), *Processing Perspectives on Task Performance*, John Benjamins, Amsterdam, pp.63-94.
5. Bui, G., & Huang, Z. (2016), "L2 Fluency as Influenced by Content Familiarity and Planning: Performance, Measurement, and Pedagogy", *Language Teaching Research*, Vol. 22, Issue 1, pp.94-114.
6. Bygate, M., & Samuda, V. (2005), "Integrative Planning through the Use of Task Repetition", in R. Ellis (ed.), *Planning and Task Performance in Second Language*, John Benjamins, Amsterdam, pp.37-74.
7. De Bot, K. (1992), "A Bilingual Production Model: Levelt's 'Speaking' Model Adapted", *Applied Linguistics*, No. 13, pp.1-24.
8. De Bot, K. (2008), "Introduction: Second Language Acquisition as a Dynamic Process", *The Modern Language Journal*, No. 92(2), pp.166-178.
9. De Bot, K., & Schreuder, R. (1993), "Word Production and the Bilingual Lexicon", in R. Schreuder & B. Weltens (eds.), *The Bilingual Lexicon*, Benjamins, pp.191-214.
10. De Bot, K., Schmid, M. S., & Lowie, W. (2011), *Modeling Bilingualism from Structure to Chaos*, John Benjamins, Amsterdam.
11. De Jong, N., & Perfetti, C. A. (2011), "Fluency Training in the ESL Classroom: An Experimental Study of Fluency Development and Proceduralization", *Language Learning*, No. 61(2), pp.533-568.
12. De Jong, N. H., Steinel, M. P., Florijn, A. F., Schoonen, R., & Hulstijn, J. H. (2013), "Linguistic Skills and Speaking Fluency in a Second Language", *Applied Psycholinguistics*, No. 33, pp.1-24.
13. Dell, G. S. (1986), "A Spreading Activation Theory of Retrieval in Sentence Production", *Psychological Review*, No. 93, pp.283-321.
14. Derwing, T. M., Rossiter, M. J., Munro, M. J., & Thomson, R. I. (2004), "Second Language Fluency: Judgments on Different Tasks", *Language Learning*, No. 54(4), pp.655-680.
15. Ellis, R. (2009), "The Differential Effects of Three Types of Task Planning on the Fluency, Complexity, and Accuracy in L2 Oral Production", *Applied Linguistics*, No. 30, pp.474-509.
16. Hilton, H. (2008), "The Link between Vocabulary Knowledge and Spoken L2 Fluency", *Language Learning*, No. 36(2), pp.153-166.
17. Koizumi, R, & Yo, I. (2013), "Vocabulary Knowledge and Speaking Proficiency among Second Language Learners from Novice to Intermediate Levels", *Journal of Language Teaching & Research*, No. 4, pp.900-913.
18. Kormos, J. (2006), *Speech Production and Second Language Acquisition*, Lawrence Erlbaum Associates.

19. Kormos, J. (2011), "Speech Production and the Cognition Hypothesis", in P. Robinson (ed.), *L2 Task Complexity: Researching the Cognition Hypothesis of Language Learning and Performance*, John Benjamins, Amsterdam, pp.39-60.
20. La Heij, W. (2005), "Selection Processes in Monolingual and Bilingual Lexical Access", in J. Kroll & A. M. B. de Groot (eds.), *Handbook of Bilingualism, Psycholinguistic Approaches*, Oxford University Press.
21. Lambert, C., Aubrey, S., & Leeming, P. (2020), "Task Preparation and Second Language Speech Production", *TESOL Quarterly*, https://doi.org/10.1002/tesq.598.
22. Lambert, C., Kormos, J., & Minn, D. (2017), "Task Repetition and Second Language Speech Processing", *Studies in Second Language Acquisition*, No. 39(1), pp.167-196, doi:10.1017/S0272263116000085.
23. Larsen-Freeman, D., Schmid, M. S., & Lowie, W. (2011), "From Structure to Chaos: Twenty Years of Modeling Bilingualism", in Schmid", M. S. & Lowie, W. (eds.), *Modeling Bilingualism: From Structure to Chaos*, Benjamins.
24. Levelt, W. J. M. (1989), *Speaking: From Intention to Articulation*, MIT Press.
25. Levelt, W. J. M. (1995), "The Ability to Speak: From Intentions to Spoken Words", *European Review*, No. 3(1), pp.13-23.
26. Levelt, W. J. M. (1999), "Producing Spoken Language: A Blueprint of the Speaker", in C. Brown & P. Hagoort (eds.), *Neurocognition of Language*, Oxford University Press, pp.83-122.
27. Liu Y. (2020), "Relating Lexical Access and Second Language Speaking Performance", *Languages*, No. 5(2), p.13, https://doi.org/10.3390/languages5020013.
28. Segalowitz, N. (2010), *Cognitive Bases of Second Language Fluency*, Routledge.
29. Skehan, P. (2009), "Modelling L2 Performance: Integrating Complexity, Accuracy, Fluency and Lexis", *Applied Linguistics*, No. 30, pp.510-532, https://doi.org/10.1093/applin/amp047.
30. Uchihara, T., & Kazuya, S. (2019), "Exploring the Relationship between Productive Vocabulary Knowledge and Second Language Oral Ability", *The Language Learning Journal*, No. 47, pp.64-75.
31. Wang, Z. (2014), "On-line Time Pressure Manipulations: L2 Speaking Performance under Five Types of Planning and Repetition Conditions", in P. Skehan (ed.), *Processing Perspectives on Task Performance* John Benjamins, Amsterdam, pp.27-62.