# Language production

* 1. **Introduction**

In this lecture, it will be seen that both models of L1 production as well as accounts of L2 production can shed light on what happens with L2 performance. In particular, this chapter will aim to provide an answer to the following four questions:

* + 1. How is language produced?
		2. How does L2 production differ from L1 production?
		3. How do attention and memory mediate L2 production and development?
		4. How can engaging in production potentially lead to second language development?

In order to understand how L2 production works we first need to understand how L1 operates, and so L1 production will be considered in the first part of this chapter. Levelt’s (1989) L1 production model will be used as the main reference. Special emphasis will be placed on the areas of message generation, retrieval of lexical items, message formulation, and self-monitoring, with only a more limited consideration of articulation and speech comprehension. In the second part of the chapter, from a number of accounts of L2 production, those aspects that distinguish it from L1 production will be pointed out, and lexical access, grammatical encoding, and self-monitoring will receive special attention. After that, the underlying constructs of attention and memory will be discussed. Regarding attention, the ideas of limited resources, selection, and capacity will be outlined, and the aspects most directly related to production will be underlined. The architecture and processes of memory will be presented, and the distinction between working memory (WM) and long-term memory (LTM) will be established. Finally, theories that link production to language learning will be outlined.

# Models of language production

In the past 30 years, the interest in language production has given rise to a number of psycholinguistic models that have tried to account for how language goes from ‘mind to mouth’. More specifically, psycholinguistic models of language production have tried to provide an explanation for the efficiency and accuracy of the system. Hence, they have tried to discover how an average speaker can produce language at a rate of 2 to 3 words per second, that is, 120 -200 words per minute, and with the very low rate of errors of approximately 1 error every 1000 words. Although most models agree that there exist distinct processing levels responsible for conceptually generating, encoding or formulating, and articulating messages, they differ considerably on how they explain the characteristics of such processes as well as the relations among them. In the last three decades, the main divide has been between modular and non-modular models of language production. Researchers embracing modular models (Garrett, 1984, 2000; Laver, 1980; Levelt’s, 1989, 1993; Levelt et al. 1999) have postulated the existence of a number of encapsulated, specialist modules or processes through which production proceeds, without interaction existing among them. In this type of models information flows unidirectionally, that is, from one component or module to the next without the possibility of feedback. For instance, the process responsible for generating messages at a conceptual level provides information to the next component which is responsible for linguistically encoding them, but this latter process, the formulator, does not send any information back to the conceptualizer, and neither does any other component.

 Another characteristic of these models is that they suggest that the information that flows from one component to the next one is the minimal necessary information, and hence information from other processing levels is simply not transmitted. On the other hand, non-modular accounts of L1 production (Dell, 1986; Kempen & Vosse, 1989; MacKay, 1987, 1992; Trueswell, Tanenhaus & Garnsey, 1994; Vigliocco & Hartsuiker, 2002) have questioned the information encapsulation and lack of interaction among components. They have advocated more flexible models in which information can flow in two directions (e.g. from the message generator to the message formulator and back) and where the input to one level can be information converging from different levels (e.g. the selection of a lexical item may be informed by both the conceptualizer, a process that precedes lexical selection, and by the processes responsible for building syntactic frames, which is supposed to be a later process ).

 Levelt’s (1989, 1993; Levelt et al. 1999) model of L1 production is used to help explain the effects of manipulating Task Complexity on L2 learners’ production. There are three reasons for choosing Levelt’s model: firstly, Levelt’s has been the most widely accepted and influential model in L2 production research, and therefore its use in this study will permit establishing comparisons to explanations and findings in other studies. Some examples of studies that have used Levelt’s model in the L2 context are Izumi’s (2003) attempt to provide a psycholinguistic rationale for the Output Hypothesis, De Bot’s (1992) and Poulisse’s (1997) account of language production in bilinguals; De Bot et al’s (1997) explanation of second language vocabulary acquisition; Poulisse and Bongaerts’ (1994) theory of L2 lexical access; and Yuan and Ellis (2003) application of the model to the explanation of the effects of pre-task and on-line planning time on production. Secondly, it is believed that Levelt’s production model, which is based on a long tradition of psycholinguistic research and on robust empirical findings, is relevant to this study because it complements the explanation of other processes which mediate language processing such as attention and memory. Levelt’s model is based on findings that have primarily been the result of the study of speech errors (e.g. tip-of-tongue phenomenon or word substitution) in both normal speakers and speakers with language pathologies (e.g. anomia, which is a kind of aphasic disturbance in which speakers have difficulties retrieving a word). In the third place, his model is in fact a further development of other proposals and it integrates specific explanations from them (e.g. Dell’s (1986) spreading activation model of lexical access, which will be further discussed in the next section), two features that enhance its explanatory power.

# Levelt’s model of L1 production

Figure 1 on the next page shows a schematic representation of the different processing components involved in spoken language use as suggested by Levelt (1989, 1993). The speech production system advanced by Levelt (1989, 1993) consists of a number of autonomous components which are responsible for different aspects of speech production. These components include: the conceptualizer, a component that is responsible for generating and monitoring messages; the formulator, in charge of giving grammatical and phonological shape to messages and which feed son the lexicon; the articulator, which specializes in the motor execution of the message; an audition or acoustic-phonetic processor, which transforms the acoustic signal into phonetic representations; and the speech comprehension system, which permits the parsing or processing of both self-generated as well as other-generated messages.



*Figure 1.* Levelt’s (1993, p. 2) model of language production.

# Message conceptualization

The first component in Levelt’s (1989, 1993) production system is the conceptualizer. This component is responsible for generating the communicative intention3 and for encoding it into some kind of coherent conceptual plan. In addition, the conceptualizer monitors what is about to be said as well as what has been said and how. In order to generate a message, declarative knowledge is accessed. Declarative knowledge includes encyclopedic knowledge (about the person’s general experience of the world), knowledge about the situation (e.g. the interlocutor/s and the communicative context, among others), as well as information about the discourse record, that is, what has already been said. Levelt distinguishes two stages in message planning: macroplanning and microplanning. Macroplanning consists of retrieving information to express the subgoals into which the overall communicative goal has been elaborated. In other words, it involves generating speech act intentions, like to narrate an event or express an opinion. In Levelt’s terms (1993, p. 3): “The speaker’s planning of a speech act, his selection of information to be expressed, and his linearization of that information are called ‘macroplanning’”. Microplanning divides that information into smaller conceptual ‘chunks’ which are given the correct propositional shape and informational perspective. For instance, the narration of a small event may be realized by a statement which can be presented in different ways (e.g. ‘the man gave the woman

Within Levelt’s model, ‘intention’ should be interpreted as ‘willingness to execute a speech plan’. Levelt (1989, p. 59) is not concerned with where intentions, in their more general sense, come from. He restricts his discussion to communicative intentions, which underlie speech acts (Austin, 1962; Searle, 1969, 1979). For a broader, philosophical definition of ‘intention’ see Dennett, D. (1987). the money” or the “woman was given the money by the man”). As noted by Poulisse (1997, p. 221), how exactly chunking takes place has not been clearly established yet. The product of macro and microplanning is what Levelt refers to as the preverbal plan, that is, an organized conceptual structure which is not yet linguistic and which constitutes the specific input that the next processing component, the formulator, will work on.

The distinction between macro and microplanning will be of special interest to us when we discuss how increasing the cognitive complexity of language learning tasks can have specific consequences for message conceptualization. It will be seen that increasing complexity of oral tasks along certain dimensions (e.g. the reasoning demands imposed by the task or the degree of displaced, past time reference) can force changes in macroplanning. These changes have a direct consequence for microplanning of the form of the utterance. For example, performing a task which has been manipulated along its displaced past time reference (i.e. from the Here- and-Now to the There-and-Then) forces microplanning of regular and irregular past tense inflections (Robinson, personal communication). The effects of manipulating cognitive complexity on macro and microplanning will also have consequences for lexical variety and complexity.

It is important to briefly highlight at this point the fact that conceptualizing the message requires attentional control. That means that the different types of information needed to express the intention have to be attended to in order for them to be retrieved from long-term memory (LTM) and instantiated into working memory (WM)4, a task which is supposed to take up memory resources.

# Message formulation

In the next component in the production system, the formulator, the propositionally organized preverbal plan activates the items in the lexicon that best correspond to the different chunks of the intended message that will, in turn, be responsible for transforming it into a linguistic structure*.* In Levelt’s model, as well as in several other models (e.g. Garrett, 1975, 2000; Kempen & Huijbers, 1983), grammatical and phonological encoding are lexically driven. For grammatical encoding to take place, both lexical access procedures and syntactic procedures are applied. In the lexicon, each lexical item is specified for semantic and syntactic information (lemmas), and morphological and phonological information (lexemes).

From a number of connectionist proposals as to how lexical access takes place (Anderson, 1983; Dell 1986; MacKay, 1987; Rumelhart et al., 1986), Levelt presents Dell’s (1986) spreading activation theory as the most promising one to account for how lexical access takes place during real time performance. In brief, a chunk in the preverbal plan activates a number of lemmas in the lexicon. The lemmas which receive the highest activation because their semantic specifications match the concepts in the preverbal plan will be selected.

For example, if a speaker wants to produce the sentence ‘The man gave the woman the money”, out of 30,000 words average speakers have active in their lexicon the four content words ‘man’, ‘give’, ‘woman’ and ‘money’ will receive the highest activation because they best match the pre-verbal plan. This does not mean that other items do not get activated. Together with ‘man’, other entries which share similar conceptual specifications get activated, but it is ‘man’ that gets the highest activation6 (See Figure 2 below).

person

**man**

woman

child

***Figure 2.* Spreading activation and selection of a lexical item.**

When a lemma is retrieved because it matches part of the preverbal message, its syntactic properties become available and they trigger syntactic building procedures. For example, the entry for ‘give’ will contain its conceptual specification and conceptual arguments7, the syntactic category (verb), the grammatical functions it requires (subject, direct and indirect objects), its relations to verbal complements (none in this case), a lexical pointer8 which points to a specific form address, and a number of diacritic parameters such as tense, mood, aspect, person, and number. The procedural knowledge stored in the grammatical encoder which is activated by the syntactic information in lemmas works to build the syntactic structure of the sentence, which Levelt refers to as surface structure9. At this point, however, the specific forms of the different elements are not fully specified yet. What we have is a string of lemmas which have been organized into phrases and subphrases according to their semantic and syntactic specifications.

The lexical pointer specified in the lemma then triggers the phonological encoding process which results in the selection of specific morphological and phonological forms (See Figure 3 on the following page). For example, if the intention of the speaker is to express ‘the man gave the woman the money ‘, ‘give’ and ‘gave’ among other lemmas will be activated but ‘gave’ will receive the highest activation because its diacritic parameter ‘tense’ matches the lemma.



Subsequently, by means of a series of phonological procedures which will draw on the syllabary, the form information of each lexical item will be further specified. Briefly put, a series of phonological segments are activated, and a phonological word is produced, which leads to the generation of the phonetic-articulatory plan. The output of the formulator is a phonetic plan or articulatory plan which is ready for articulation10. Levelt refers to this articulatory plan as internal speech, as opposed to already articulated overt speech.

In contrast with conceptual generation, which we saw is an attention and memory demanding process, message formulation of adult monolinguals is seen as an automatic process that will ensure a relatively fast flow of speech at the rate of 3 to 5 words per second (300 words per minute).

# Articulation

The articulator is the next component in the speech production system. Articulation is the motor execution of the phonetic plan, and it involves the respiratory, the laryngeal, and the supralaryngeal systems*.* Briefly put, as the phonetic plan is being generated, its bits are temporarily stored in an ‘articulatory buffer’. This buffered information will trigger the ‘unpacking’ of motor commands which will finally cause articulation of the message. It is this buffered speech which speakers subjectively experience as internal speech. The outcome of articulation is overt speech. We will not detail this process any further since it is beyond the main concerns in this dissertation.

# Audition, speech comprehension, and monitoring

As previously stated, the conceptualizer is in charge of both generating messages and monitoring the whole process of production. In Levelt’s account, speakers make use of their speech comprehension system to listen to and process their own speech in exactly the same way they listen to and process the speech of others11. The difference is that speakers have access to both their internal and their overt speech12. In the case of internal speech, at an early stage the preverbal plan can be checked against the speaker’s intention. Later on the process, the articulatory plan is representable in working memory where it can be checked. In this way the speaker can detect problems before he or she has articulated an incorrect item. As for overt speech, the audition component of the system recognizes the articulated words, and the speech comprehension system will retrieve their meaning. Hence, learners monitor both the meaning and the well-formedness of their productions. When a problem is detected, several options are available, such as simply ignoring the problem, revising the preverbal message, or generating a new message. As will be detailed further in Section 1.4.3, this will depend on the nature of the problem as well as on other contextual factors.

Summing up what we have so far, if a speaker wants to produce the utterance “the man gave the woman the money”, he or she will first pay attention to what he or she wishes to say. He or she will then select the information he or she needs from his or her encyclopedic knowledge, and by considering the communicative situation and what has been said so far. As chunks of the intended conceptual message are decided on, and if everything goes well, he or she will activate and select the appropriate lemmas which will construct the surface structure of the utterance and that will point to the most appropriate word forms. Once each lemma is given a morphological and phonological form, articulation will begin and overt speech will take place. Throughout the whole process, the conceptualizer will supervise the message by checking the pre-verbal plan against the intention, the pre-articulatory plan against the conceptual plan, and the already uttered message against what was intended.