Chapter I

Language production

1.1 Introduction

In this chapter 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:

i) How is language produced?

ii) How does L2 production differ from L1 production?

iii) How do attention and memory mediate L2 production and development?

iv) 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.

1.2 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)[1].

In this study, 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

[1] See Section 1.3.5 for specific examples.

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.

Throughout this chapter, we will see that some crucial assumptions of Levelt's model, such as the minimal input received by each component or the unidirectional flow of information which were mentioned before have been questioned by other researchers. When appropriate, then, reference to alternative accounts of some specific processes will be made.

1.3 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 feeds on the lexicon[2]; 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.

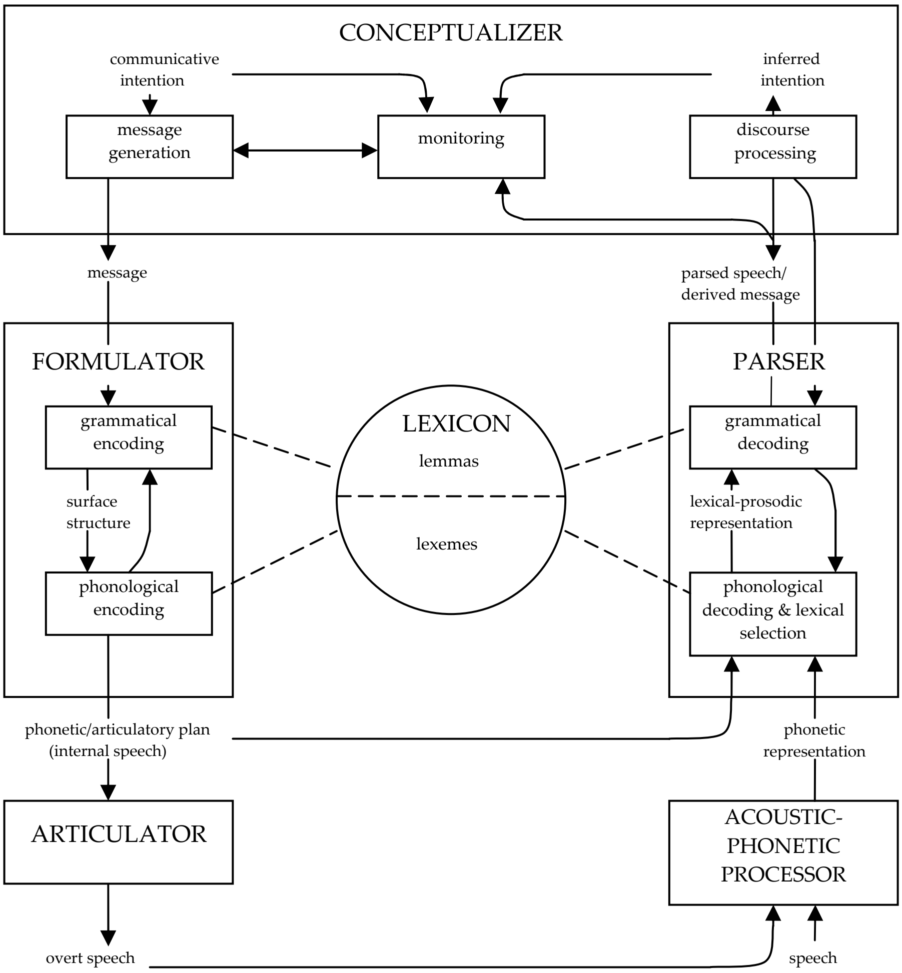

Figure 1. Levelt's (1993, p. 2) model of language production.

[2] In Levelt's terms (1993, p. 4), the mental lexicon "is the repository of knowledge about words in one's own language".

1.3.1 Message conceptualization

The first component in Levelt's (1989, 1993) production system is the conceptualizer. This component is responsible for generating the communicative intention[3] 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

[3] 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. This issue will be specifically discussed in Section 8.4.2.

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. The issues of attention and memory will be dealt with in greater detail later on in this dissertation (See Sections 1.5.1 and 1.5.2).

1.3.2 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

[4] As it will be seen later on, working memory is that part of long-term memory which in a current state of activation.

receive the highest activation because their semantic specifications match the concepts in the preverbal plan will be selected[5].

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 activation[6] (See Figure 2 below).

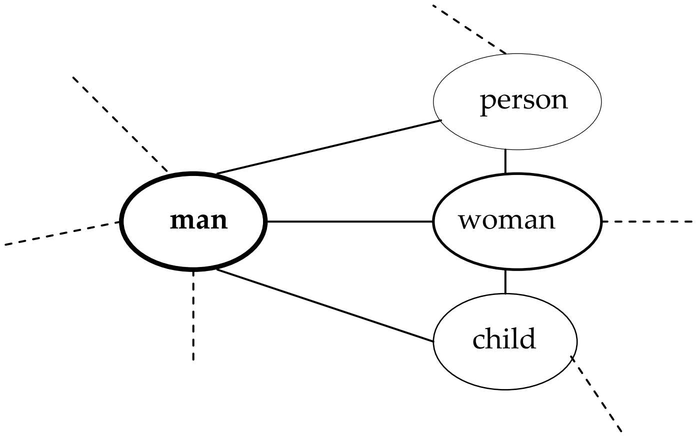

Figure 2. Spreading activation and selection of a lexical item.

[5] As Levelt (1993, p. 4) suggests, although a number of lexical access theories exist, they are still "seriously inadequate and underspecified".

[6] The issue of how much overlap must exist between the concept specification and the selected item is still the subject of much debate. De Bot et al. (1997, p. 312) suggests that selection is determined, among other factors, by the demands of the conversational setting.

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 arguments[7], 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 pointer[8] 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 structure[9]. 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.

[7] For example, 'give' will specify that the agent it requires will act as a subject in the surface structure. This is what Levelt refers to as argument-to-function mapping.

[8] As Levelt (1989, p. 165) puts it: "A lemma's lexical pointer indicates an address, where the corresponding word-form information is stored…the pointer to a form address may be indexed with various features that will affect the word-form retrieved. We will call them diacritic features."

[9] Some mechanisms of grammatical encoding have been proposed. For example, the Incremental Procedural Grammar model by Kempen and Hoenkamp (1987) or the competition model by MacWhinney and Bates (1978). Their detailed consideration is beyond the scope of this work.

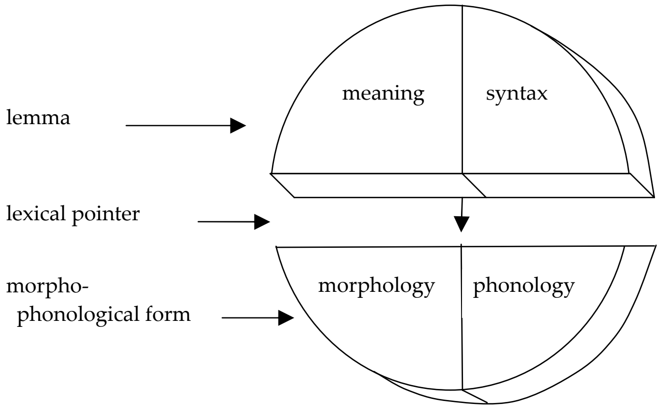

Figure 3. Levelt's (1989, p. 188) representation of a lexical entry.

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 articulation[10]. Levelt refers to this articulatory plan as internal speech, as opposed to already articulated overt speech.

[10] There is a debate about whether grammatical information and phonological info are organized in a modular (Levelt, 1989) or in an interactive fashion (Dell, 1986). In the modular fashion, semantically and syntactically specified lemmas are accessed first, and then phonological access follows. In the interactive proposal, there is overlap, and it is suggested that lemmas are more activated at the beginning whereas, phonological info is more activated towards the end of lexical access, but activation takes place in parallel.

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). As we will see later on, however, for L2 speakers, automaticity is not a given feature of language production.

1.3.3 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.

1.3.4 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 others[11]. The difference is that speakers have access to both their internal and their overt speech[12]. 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

[11] Although there is robust evidence that supports this statement, it has been suggested that this 'conscious' kind of feedback is not the only kind of feedback in the production system. Vigliocco and Hartsuiker (2002, p. 466) suggest the existence of an inner monitoring mechanism (outside the speech comprehension system) which operates locally between components and without the speaker's awareness (See example in Section 1.2.5, second paragraph).

[12] In contrast to this view, Vigliocco and Harsuiker (2002, p. 467) suggest that: "perception can monitor either our overt speech or our silent speech, but not both of them at the same time."

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.

This description may give the impression that production happens in a linear fashion, by constructing the elements one after the other, but as will be seen in the next section, this would make it impossible to achieve an acceptable rate of speech.

1.3.5 Assumptions about components

Before we go on to analyze how Levelt's model can be adapted to explain L2 production, it is important to mention that this particular architectural representation of the speech production system (see Figure 1 on page 26) makes a number of assumptions regarding the characteristics of its components.

Firstly, the components in Level's model are specialist processing systems which, although they may need as input the output of other components, do not need to share information with other components to carry out their job. For example, the formulator needs the concepts provided by the conceptualizer (i.e.

the pre-verbal plan), but its processing is completely independent of any other components, such as the articulator or the acoustic-phonetic processor. Not only that, there is no interference among components in the sense that the way one component operates is not affected by the output of other components, and they only take a specific kind of input and not the input which is specific to other components. Each component contains its own procedural knowledge and, as will be outlined later, the only possible feedback is that provided by internal or overt speech to the conceptualizer, by means of monitoring. Information is only fed forward (e.g. from the conceptualizer to the formulator, but not from the formulator to the conceptualizer), and there is no interaction between components. These assumptions of Levelt's modular architecture have been challenged by Vigliocco and Hartsuiker (2002), among other authors. They reject Levelt's idea that only the 'minimal' necessary information flows among the different levels. They also oppose Levelt's idea that there exists a unidirectional flow of information, with no possibility of feedback. In their view, 'maximal' input may imply that, for example, in the case of lemma retrieval and phrasal construction, phrasal construction can occur before lemma selection, not only for the lemma that will be selected but also for competitor lemmas that are highly activated (See Figure 4 on the following page). In the case of information flow, sublexical units, for instance, may inform lexical units in such a way that the level of activation of the lexical target increases, its selection is ensured, and errors are avoided. In other words, feedback between components contributes directly to the efficiency and accuracy of production.

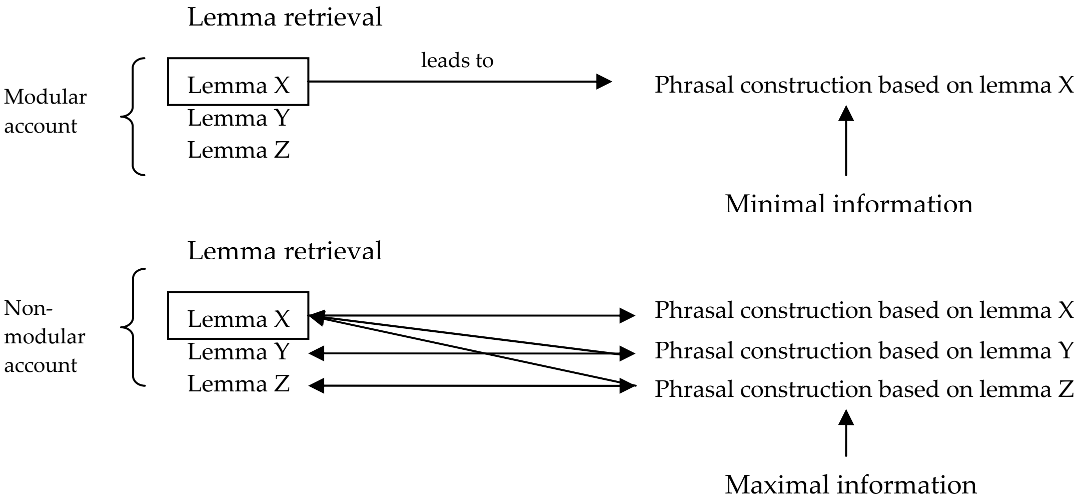

Figure 4. Representation of modular and non-modular accounts of language production. The case of lemma retrieval and phrasal construction.

Secondly, the different components can process their specific types of information simultaneously or in parallel. Normal language productions require different components to be working simultaneously for speech to flow at an acceptable rate. In Levelt's model, processing operates both serially and in parallel. That explains why connectionist models of spreading activation are so useful to explain what happens in production, since they allow for parallel processing which is necessary to account for real time production.

Thirdly, related to the idea of parallel processing is the idea that production is incremental. When one component has processed its specific input and passes it on to the next component (e.g. from the conceptualizer to the formulator), it can start processing a different chunk of input while its output is being processed in the next component. For example, in a sentence like 'the man gave the woman the money', after preverbal plan for ‘man’ has finished and starts being processed by the formulator, the conceptualizer can start working simultaneously on the plan of ‘gave’ or ‘woman’ without waiting for the outcome of the formulation of ‘man’. In Levelt’s (1993, p. 12) words, “though all components work in parallel, they work on different bits and pieces” of the message.

In the fourth place, at least some parts of the processing system work automatically. In Levelt’s model, the conceptualizer requires executive control, that is, conscious allocation of attentional and memory resources for both message generation and monitoring. As Levelt (1989, p. 21) states: “Human controlled processing tends to be serial in nature, and is therefore slow.” Conceptualizing a message requires a number of steps, such as constructing an internal representation (e.g. from perceived input or from inferences and deductions from stored knowledge), selecting the information to be communicated, breaking it into smaller chunks, and organizing them in a linear fashion (Guhe, 2003), a process which shares processing resources with monitoring. Conversely, grammatical and phonological encoding are assumed to be automatic, which means that they do not require attention because they are single-step processes. In Poulisse’s (1997, p. 204) words: “the grammatical and phonological encoding of a message, including lexical articulation, are usually automatic. With some exceptions, for example, in the case of very infrequent words, these processes are executed without conscious awareness and do not share their processing resources with other processes.” Following De Bot

(1996), it can be concluded that parallel processing, incremental production, and automaticity allow for the speedy production of language in real time.

So far we have attempted to provide an answer to the first question we posed in Section 1.1., and thus we have tried to explain how language is produced. Before we turn to the second question, though, it is worth mentioning that some of the components in Levelt's model described so far will be especially relevant to the explanation of the findings in this study regarding the effects of Task Complexity on the three dimensions of L2 production, that is, fluency, complexity, and accuracy. Hence, of particular relevance to us are the processes involved in conceptualizing the message, since macro and microplanning have consequences for the dimension of linguistic complexity, both structural and lexical, during L2 production. The monitoring function which is also performed by the conceptualizer, and which will be further detailed when we discuss L2 production, will be of particular interest when we look at how task complexity affects the dimension of accuracy.

1.4 L2 production

While it can be argued that most aspects of L2 production can be explained by models of L1 production, there are some particularities of L2 speech, especially that of unbalanced bilinguals or L2 learners, that require a different kind of explanation. Poulisse (1997) suggests that there are three major differences between the two types of production which have to do with the size of the lexicon and the specification of its items, the degree of automaticity with which processing proceeds, and the presence of traces of the L1 in L2 production.

Firstly, L2 knowledge is not as complete as L1 knowledge. L2 speakers may find difficulties in expressing some concepts because, for example, they cannot find the right words to express them. Furthermore, some lexical items in the lexicon are not fully specified for their semantic, syntactic, morphological, and phonological information, which often leads to errors. Also, the relationships between different lexical items may not be fully specified yet. When an L2 speaker has trouble retrieving a lexical item, a number of compensatory strategies (Poulisse, 1990) or problem-solving mechanisms (Dörnyei & Kormos, 1998) are normally applied (See Section 1.4.3). Additionally, grammar knowledge of the L2 is underdeveloped, which may lead L2 speakers to avoiding certain L2 grammatical structures or to the production of ungrammatical sentences.

Secondly, certain aspects of processing in the L2 production system lack automaticity. While message formulation is thought to be quite automatic in the case of the L1, lexical retrieval and encoding may require serial processing for L2 speakers. As Poulisse (1997, p. 208) suggests, this mainly happens at the morphophonological and articulatory levels. In general, L2 production tends to be more hesitant and the rate of speech slower, and the degree of automaticity may vary depending on the L2 speaker's proficiency. Poulisse (1997) also notes that L2 speakers produce twice as many slips of the tongue in the L2 than in the L1.

In the third place, L2 production carries traces of the L1. L1 use during L2 production can be the result of intentional code-switching (Appel & Muysken, 1987; Faerch & Kasper, 1983; Grosjean, 1982; Poulisse, 1990;) or the consequence of unintentional performance switches (Kellerman & Sharwood-Smith, 1986; Poulisse & Bongaerts, 1994). In the case of intentional code-switching, the possible reasons behind it are multiple: one particular lexical item may be missing; there may be greater availability of L1 words; the speaker may wish to emphasize his or her identity; a change of subjects is intended; the speaker wants to specify a particular addressee, to express an emotion, or simply to mark asides from the ongoing discourse (Poulisse & Bongaerts, 1994, p. 36). Unintentional switches are commonly referred to as transfer or cross-linguistic interference[13] (Faerch & Kasper, 1986; Kellerman & Sharwood-Smith, 1986). Furthermore, the level of proficiency may affect both intentional and unintentional code-switching. Low proficiency learners may willingly resort to L1 words when they are communicating in the L2. Low level learners display more instances of involuntary transfer when speaking the L2 than high proficiency learners.

In fact, models of L2 production spring from the need to account for how languages can be kept separate and how they can get mixed. This study adopts the proposals of L2 production advanced by Poulisse and Bongaerts (1994) and Poulisse (1997) as the most satisfactory ones to account for the processes involved in with L2 performance.

[13] See Section 1.4.3.1 that elaborates on code-switching as well as on other problem-solving mechanisms in the L2.

1.4.1 Conceptualization in the L2

As has been seen, Levelt (1989) suggested two main functions of the conceptualizer. Its job is, firstly, to generate messages and, secondly, to monitor the whole speech production system. This section describes the former function while the latter will be analyzed in detail in Section 1.4.3.

As suggested by Kroll (1993), conceptual representations are shared by L1 and L2, whereas lexical representations are not. In contrast to proposals that have postulated the existence of specific subsets for the lexicon (Albert and Obler, 1978; Lipski, 1978; McNamara & Kushnir, 1971; see Kroll, 1993, for a review), one for each language, Poulisse and Bongaerts (1994) sustain that there is one single network or lexicon where L1 and L2 lexical items co-exist. Proposals that have advocated the existence of one single lexicon have also differed in their explanations as to how code-switching takes place. For example, De Bot (1992) affirmed that two pre-verbal plans were generated, one in each language. Hence, the occasional code switching episodes (whether intentional or not) in L2 speech were explained by the availability of two plans. In Poulisse and Bongaerts' account, the conceptualizer is in charge of generating the preverbal plan that specifies both the conceptual information and the language to be used. Concepts are generated and tagged for L2, and the conceptual information and the language label together activate the lemmas which contain the appropriate meaning and language (Poulisse, 1997, p. 216). Apart from the specification of language in the preverbal plan, conceptualization works exactly the same as for L1 production.

In the example we used for L1 production, the macroplan for an utterance like "the man gave the woman the money" would be the same whether the speakers wished to speak in the L1 or in the L2. If speakers wish to speak the L2, the concepts that they want to express will already indicate that lemmas and forms will be selected in the L2.

1.4.2 Formulation in the L2

So when a learner intends to speak in the L2, the different concepts in the preverbal plan will contain the specification [+ L2]**.** Spreading activation procedures operate in the same way as was described for the L1. Elaborating on the example that has been used so far, when the L2 speaker wishes to say 'man' several lexical items which are similar in meaning will be activated, including the L1 lemma (Poulisse & Bongaerts, 1994, p. 216) (See Figure 5 below).

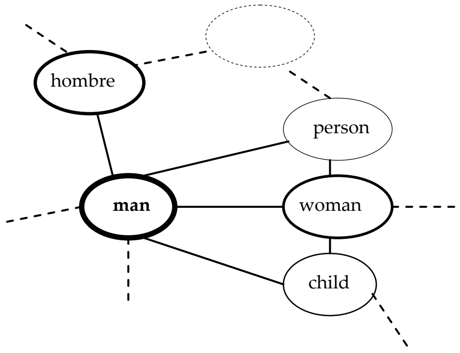

Figure 5. Spreading activation and selection of a lexical item in the L2.

This is because lexical items which belong to the two different languages are related to common conceptual nodes. For low proficiency learners, it is not uncommon that the L1 lemma receives more activation than the L2 ones even if the intention is to use the L2. This is explained by the fact that L1 lemmas, especially function words, are usually more active in the L1 than in the L2.[14]

For Poulisse and Bongaerts (1994), morphological encoding is language specific, that is, the morphemes of a message that has been planned in the L2 will come from the L2. Also in their model there is one single store of sounds and pitch patterns that is drawn on during phonological encoding. As with the lexicon, phonological errors can be explained by the fact the store is incomplete or because its items are not sufficiently specified.

Again, the issues of lexical access and formulation will be crucial for our understanding and interpretation of the findings in this study. It will be shown that the manipulation of the cognitive demands of tasks has specific consequences for how quickly or slowly learners produce, how structurally and lexically rich and varied production becomes, and what levels of accuracy learners achieve.

[14] This statement is based on experimental evidence that has analyzed 'blends'. A blend is the result of mixing two lexemes that share the same conceptual representation in two different languages. An example is provided by Poulisse (1994) of the blend 'cwame', which is a mixture of Dutch 'Kwam' and English 'came'. 'Blend', as used here, should be distinguished from 'blend' as a process in L1 morphology.

1.4.3 Self-monitoring in the L2

As we saw in Section 1.2.4, the conceptualizer is in charge of monitoring the whole process of production, an operation that requires conscious attention. More specifically, there are three monitor loops in the production system. One checks the preverbal plan against intentions. Another one checks the internal, articulatory plan against the overall plan, and a third one that monitors overt speech by means of the acoustic-phonetic processor.

When communicating, speakers find problems in their processing of messages a lot less often in the L1 than in the L2. This section is concerned with the problems L2 speakers are faced with and the kind of problem-solving mechanisms (PSM) they apply. The work of Poulisse (1993), Dörnyei and Scott (1997), Dörnyei and Kormos (1998), and Kormos (1999) will provide us with a general picture of the kind of problems speakers run into when communicating in the L2 and the array of solutions they apply[15] (See Tables 1, 2, 3, and 4 in the following pages for a summary of problem-solving mechanisms with examples).

Dörnyei and Kormos (1998) draw on Levelt's model to analyze and classify the types of communicative problems that L2 speakers are faced with, which they then link to the variety of taxonomies of communication strategies existing in the research literature (Bialystok, 1990; Faerch & Kasper, 1983; Tarone, 1977; Paribakht, 1986; Poulisse, 1993; Willems, 1987). By drawing on such rich literature, they

[15] It is important to mention that the term 'problem-solving mechanism' includes the concept of 'communication strategy'. Poulisse (1993) and Dörnyei and Scott (1997) talk about communication strategy while Dörnyei and Kormos (1998) use the term 'problem-solving mechanism'.

propose a framework that suggests a problem-solving mechanism for each type of problem. The main categories of problems have to do with:

i) resource deficits (e.g. because of an incomplete lexicon or insufficient morphological, or phonological specification);

ii) processing time pressure;

iii) perceived trouble in own output;

iv) perceived problems with the interlocutor's output[16].

The problems related to resource deficits and the kind of mechanisms that are applied to solve them will receive special consideration in the following sections.

1.4.3.1 Resource deficit problem-solving mechanisms

Dörnyei and Kormos (1998) suggest that learners can have difficulty retrieving lexical items from their incomplete L2 lexicon as well as grammatically and phonologically encoding their messages because the items in the lexicon are not sufficiently specified. When a lexical item cannot be retrieved three main options are available, L2 speakers may: i) abandon their macro-plan; ii) change their macro-plan by reducing it; iii) or change it by replacing it with a new message (See 'content reduction' in Table 1 on page 48). As Dörnyei and Kormos (1998, p. 362) point out,

[16] In this study we focus only on problem-solving mechanisms involved in monological production, and, therefore, interactive problem-solving mechanisms like confirmation requests and comprehension checks are not included.

however, this is a way of 'getting over the problem situation' rather than a way of solving the problem.

An alternative mechanism is to maintain the macro-plan and just modify the preverbal message. Poulisse (1993), who speaks about compensatory strategies rather than problem-solving mechanisms, suggests that there are three types of strategies. Firstly, a substitution strategy in which one lemma in the preverbal message may be changed or omitted, and therefore replaced by an alternative lemma. Some specific substitution strategies would be to code-switch (i.e. use an L1 term), to make use of an approximation (e.g. use a word similar in meaning like 'sky' instead of 'ceiling' or a superordinate like 'person' instead of 'man'), to utilize an all-purpose word (e.g. 'stuff' or 'thing'), or to completely omit the word (See "Substitution" in Table 1 on the following page).

Secondly, the speaker can modify the conceptual information of the lemma and apply L1 or L2 grammatical and phonological encoding processes, which usually have a wrong word as an outcome. Some specific examples of substitution plus strategy are foreignizing (e.g. 'straighten', with an English pronunciation, taken from Spanish 'estrechar' in 'estrechar las manos' which corresponds to English 'shake hands'), word coinage (e.g. 'examinates'), or literal translation (e.g. the use of 'discussion' when meaning 'argument', a false friend for Catalan and Spanish speakers) (See "Substitution plus" in Table 1 on the following page).

Table 1

Summary of problem-solving mechanisms (PSM) related to L2 lexical resource deficits based on Poulisse (1993), Dörnyei and Kormos (1998), Dörnyei and Scott (1997).

PSM related to lexical deficit	**Description**	**Example and source**
Content reduction: message abandonment	Leaving message unfinished because of difficulties with language.	“she's like hiding behind the sofa well not behind like well eh in the fourth one…” Own corpus: learner can’t come up with the exact description, leaves message unfinished, and goes on with the narration.
Content reduction: message reduction	Avoiding problematic structures or topics because of lack of linguistic resources	“he is responsible…for the…for the cleanness of the house and er…he locks the door…at night and opens…it in the morning.” Dörnyei and Kormos (1998, p. 359): in retrospective account, learner said he or she wanted to say more things but did not know how to say them so he or she made the message shorter.
Content reduction: message replacement	Replacing message with a new one because of not feeling able to execute it.	“you can…stay here until…midnight or…how do you want.” Dörnyei and Kormos (1998, p. 359): learner said he or she wanted to say ‘dawn’ or ‘morning’. He or she found the words, but he or she did not like them, and because he or she was supposed to say something, he or she came up with a new message.
Substitution: code-switching	Including L1 (or L3) words in L2 speech.	“a man who have not any any hair in in the face in the *cap*” Own corpus: learner can’t find the word ‘head’ and uses an L1 (Catalan) word.
Substitution: approximation	Using a single alternative lexical item that is similar in meaning (e.g. a superordinate).	“the man is pointing at the the *sky*” Own corpus: learner did not know the word “ceiling”.
Substitution: use of all-purpose words	Using an ‘empty’ lexical item in a context where a specific word is lacking.	The overuse of *thing* or *stuff*. Dörnyei and Kormos (1998,p .360).
Substitution: complete omission	Leaving a gap and continuing with speech.	“then…er…the sun is is…uhm sun is…and the Mickey Mouse…” Dörneyei and Kormos (1998, p. 360) report learner was not familiar with the word ‘shine’.
Substitution plus: foreignizing	Using an L1 (or L3) word by adjusting it to L2 phonology or morphology.	“they are er straight *straighten* eh their hands.” Own corpus: the learner takes Spanish “estrechar” from the expression “estrechar las manos” (shake hands) and makes it sound English.
Substitution plus: grammatical word coinage	Creating a nonexisting L2 word by applying a supposed L2 rule to an existing L2 word.	“the the doctor *examinates* hi his hair.” Own corpus: learner tries to say ‘examines’, uses English morphology but creates the wrong word.
Substitution plus: literal translation	Translating literally an L1 (or L3) lexical item to L2.	“during the *discussion* er there is there is the lady and another old old man.” Own corpus: learner meant ‘argument’ which in Spanish is ‘discusión’ and in Catalan ‘discussió’, a typical false friend.
Macro - reconceptualization: restructuring	Abandoning a verbal plan and communicating the message according to an alternative plan.	“mister Perelman is is the is a man who is eh is…com es diu? (what do you call that?)…I can’t remember…bueno (well)…Guelda is his wife.” Own corpus: learner could not produce the word “married”.
Micro-reconceptualization: circumlocution	Providing examples, illustrating, or describing the properties of the target object or action.	“there was…cómo se dice esto (what do you call that?) there was like a room eh a lot of bueno (well) a lot of mens are were waiting.” Own corpus: learner could not produce ‘waiting room’.
Micro-reconceptualization: semantic word coinage	Creating a nonexisting L2 word by compounding existing words.	‘*snowsculpture’* for ‘snowman” Dörnyei and Kormos (1998, p. .361).

Table 2

Summary of problem-solving mechanisms (PSM) related to L2 grammatical resource deficits based on Dörnyei and Kormos (1998), Dörnyei and Scott (1997).

PSM Related to grammatical deficit	**Description**	**Example and source**
Substitution	Certain grammatical specifications of the lemma are changed through transfer or overgeneralization.	"*Appears* the man and the woman is scared". Own corpus: the verb 'appear' is underspecified in L2 lexicon so L1 rule from 'aparecer' (Spanish) or 'aparèixer' (Catalan) is applied.
Reduction	Using simplified grammar in the belief that the interlocutor will be able to reconstruct the grammatical meaning from the context.	"When she er come back again". Dörnyei and Kormos (1998, p. 361): learner reported that he hesitated about which verb tense to use and stuck to the present because it was the easiest tense.

Table 3

Summary of problem-solving mechanisms (PSM) related to L2 phonological resource deficits based on Dörnyei and Kormos (1998), Dörnyei and Scott (1997).

PSM Related to phonological deficit		
Retrieval: tip-of-tongue phenomenon	When attempting to retrieve a lexical item, speaker utters a series of incomplete word forms until reaching the complete form, or uses several alternative pronunciations until one sounds right.	"In the picture there was a *cou coup* er *couple* in bed trying to sleep." Own corpus: learner utters parts of the word until the complete word is achieved. "the man is laugh [laf] er laugh [lof] laughing [laujin]". Own corpus: learners tries out several options before deciding on one.
Substitution: use of similar sounding words	Compensating for a lexical item whose form the speaker is unsure of with a word (either existing or non -existing) that sounds approximately like the target.	"his social class is very *slow*." Own corpus: learner wanted to say 'low'
Reduction: mumbling	Muttering or swallowing a word or a part of it because speaker is not sure about it.	"And uh well Mickey Mouse looks surprise or sort of XXX" Dörnyei and Kormos (1998, p. 362): they suggest that 'sort of' indicates that the unintelligible part is not a recording failure but a strategy.

Thirdly, a chunk of the preverbal message can be reconceptualized. In the case of a macro-reconceptualization this means that a series of chunks that are part of the overall plan are completely abandoned and a new plan is generated, a mechanism that Dörnyei and Kormos (1998) refer to as 'restructuring'. As for micro-reconceptualization, it involves revising a single chunk of the message by using a circumlocution or semantic word coinage (See "Macro-reconceptualization and Micro-reconceptualization" in Table 1 on page 48).

As far as problems with grammatical encoding are concerned, one of the options is to change some characteristics of the lemma in terms of form or argument structure. If, for example, the arguments of a specific lemma are not sufficiently specified, the speaker may apply the rules and argument structure of his or her L1. This is the typical case of Catalan and Spanish speakers of English who use sentences like "appears the man". The verb 'appear' in English is not sufficiently specified in their lexicon, and therefore they apply their L1 rule which establishes that 'aparèixer' in Catalan or 'aparecer' in Spanish the subject can appear after the verb. A second alternative solution is to reduce or simplify the grammar and hope that the interlocutor will be able to reconstruct the grammatical meaning from the context of the interaction (See Table 2 on previous page). A fourth option that Dörnyei and Kormos (1998) suggest is not to activate the problematic lemma. However, this option does not appear to be feasible since activation is automatic and involuntary.

Finally, for phonological encoding, again one of the options is to avoid the term that cannot be verbalized. A second one is to try out several alternatives (e.g. to verbalize 'laugh', Catalan or Spanish speakers may produce two or three alternative pronunciations [laf] [lof] [lauj]) before choosing one. Two other alternatives would be to apply articulatory substitution (e.g. use 'slow when 'low' cannot be verbalized) or 'swallow' sounds, which Dörnyei and Scott (1997) classified as 'mumbling' (See Table 3 on page 49).

1.4.3.2 Time-pressure problem-solving mechanisms

During real time performance, speakers are aware that excessively long pauses are unacceptable. When they figure that production is going to be hampered by too many problems that will impede an acceptable speech rate they may opt for a number of solutions. The first solution would involve reducing or abandoning the message that is taking so long to process. They can also use the resource-deficit strategies that were analyzed in the previous section. And they can also make use of stalling mechanisms. These include time gaining devices such as unfilled pauses or nonlexicalized pauses (e.g. such as 'uhm' or 'er'), lengthening of sounds, fillers (e.g. 'you know', 'I mean', etc.), and repetitions, which are far from uncommon in L2 speech (See Table 4 on the following page).

Table 4

Summary of problem-solving mechanisms (PSM) related to time pressure based on Dörnyei and Kormos (1998) and Dörnyei and Scott (1997).

PSM related to processing time pressure	Description	Example and source
Unfilled pauses	Speaker remains silent while thinking.	"the man is / / / is / / / angry." Own corpus: learner is silent for a few second until the word is found.
'Umming' and 'erring'	Use of nonlexicalized filled pauses like *er, uhm,* and *uh.*	"so *er* mister Roper *er* wake up". Own corpus: learner cannot remember the name of the character and uses filled pauses to plan message
Sound lengthening	Lengthening of a sound.	"suddently ([s:::adentli]) no suddenly". Own corpus: learners lengthens the "s" because of being unsure about the accuracy of the word, as shown by subsequent self-repair.
Fillers	Time-gaining devices such as *okay* or *you know* which maintain discourse when there are difficulties.	"this is the story of a man and woman who are married *okay* and they they are in a room." Own corpus: learner stalls and says "okay" probably to plan the sentence ahead and to maintain the fluent pace of the narration.
Repetitions	Repeating a word or a string of words right after they were said, or right after the interlocutor said them to gain time.	"they go *to to the to the* door." Onw Corpus: learner repeats word to gain time, either to find words or plan what to say next.

1.4.3.3 Deficient own output problem-solving mechanisms.

Probably the most exhaustive taxonomy of problem-solving mechanisms related to perceived deficiencies in one's own production, more commonly known as self-repair, is the one advanced by Kormos (1999). Her framework brings together theories about psycholinguistic processes of production, L2 research findings regarding the development of automaticity, as well as theories of consciousness, awareness, and noticing.

Kormos adopts Levelt's perceptual loop theory of L1 monitoring. In agreement with van Hest (1996), Kormos suggests that perceptual loop theory can be used to explain monitoring in the L2 with no major qualitative changes. Nonetheless, there are at least two phenomena that the perceptual loop theory does not explain. Firstly, despite having sufficient underlying L2 knowledge, a lot of errors in L2 speech are not noticed by L2 speakers. Secondly, certain types of errors go undetected. In Kormos' view, differences in monitoring are explained by the fact that monitoring needs attentional control. One explanation of this is that because attentional capacity is limited, attention used to compensate for the non-automatic processes in the L2 (e.g. grammatical and phonological encoding) limits the amount of monitoring that takes place. This explanation assumes, then, that because a lot of attentional resources are being used to encode messages both grammatically and phonologically in the L2, fewer resources are available for self-monitoring which, as consequence, is reduced. As will be seen in Section 1.4.1.3, this limited-resources view of attention is challenged by current multiple-resources accounts of attention. An additional explanation by Kormos is that attention to monitoring depends on individual differences, which display different working memory capacities among speakers (Harrington & Sawyer, 1992). Thirdly, in Kormos' view, increasing task demands deviates attention from monitoring. Again, this assertion deserves closer scrutiny since there are opposing views about how task demands affect monitoring. Skehan (1998, Skehan & Foster, 2001; Foster and Skehan, 1996; Skehan & Foster, 1997), for example, suggests that as language tasks are made more demanding by, for example, giving learners less time to plan them, they consume capacity which affects monitoring, with negative consequences for accuracy. Robinson (1995b; 2001a; 2001b, 2003a; forthcoming) agrees with Skehan that increasing task demands along planning time, as well as other 'resource-dispersing' dimensions of tasks, deviates attention from monitoring. He argues, however, that increasing task demands along other dimensions, such as the number of elements in a task, the reasoning demands they impose on speakers, and their degree of displaced, past-time reference (which he refers to as 'resource-directing' demands) actually draws learners' attention to the way they encode messages with positive consequences for accuracy (See Section 2.4.3 for a detailed discussion of these views). The study in this dissertation specifically addresses this crucial issue of how task complexity affects self-monitoring in Chapters V and VII.

Following the previous classification of repairs (Kormos, 1998; Levelt, 1983; Van Hest, 1986), Kormos distinguishes between different repairs (D-repairs), appropriateness repairs (A-repairs), and error repairs. The first type of repair is the consequence of errors with the conceptualization of the message, either because the information has not been organized properly or because it has been inadequately encoded in the preverbal plan[17]. This basic classification taken from the work of Levelt (1983) was extended by Kormos (1998) to include inappropriate information

[17] Neither Dörnyei and Kormos (1998) nor Kormos (1999) provide a clear-cut distinction between 'message replacement' and 'different information repairs". The only explanation may be found in the fact that the first type is due to a lack of vocabulary rather than to a decision to provide a different kind of information for reasons other than lack of resources. In any case, the classification of this type of repairs has to be based on the interpretation of retrospective protocol accounts.

repairs, message replacement repairs, and message abandonment repairs[18] (See "D-repairs" in Table 5 on the following page). Appropriateness repairs are meant to resolve ambiguity, achieve precision, and maintain coherence with previously used words. To these three she adds Brédart's (1991) pragmatic appropriacy repairs and repairs for 'good language' (See "A-repairs" in Table 5 on the following page). Finally, error repairs are the result of wrong formulation, and may be used to correct an inappropriate syntactic structure, a lexical problem, faulty morphology, or a phonetic error (See "Error repairs" in Table 5 on the following page). It is worth mentioning that the identification of the reasons behind different and appropriate repairs (i.e. the type of error they are correcting) require the use of retrospective protocol analysis for their identification. Conversely, overt error repairs are more easily identified and do not require introspection for their identification and classification. This latter type is the only type of self-repair the experiment in this study will be concerned with.

[18] Again this problem-solving mechanism is similar to the "message abandonment" in Table 1, only that this time it is not caused by a deficit of lexical resources but because the speaker comes up with a new idea and abandons the first one.

Table 5

Summary of problem-solving mechanisms (PSM) related to deficient own output based on Dörnyei and Kormos (1998) and Kormos (1999).

PSM related to deficient own output	Description	Example and source
D-repair: different information repair	They imply the reconceptualization of the speech plan.	"We go straight on or…we come in via red, go then straight to green." Levelt (1983, p. 51), supplied by Kormos (1999, p. 318): the speakers changes the original speech plan by encoding different information.
D-repair: inappropriate information repair	The speaker detects a problem with the information content and repairs it.	"then the man ay the *man* the woman leaves the room". Own corpus: learner detects wrong information in message and repairs it
D-repair: different order repair	The speakers changes the order of the different parts of the message.	"well we it's it's about a thousand forints…" Levelt (1983), provided by Kormos (1999) reported that learner began by answering the second of two questions he or she was asked but changed his or her mind and answered the first question first.
D-repair: message abandonment repair	Learner replaces the originally intended message by a new one.	"we have some ere r v- maybe you have vegetarians in your group. Kormos (1999, p. 380): the speaker reported that a new idea 'popped up' and abandoned the original one.
A-repair: appropriate-level-of-information repair	The speaker decides to be more precise or specific.	"then a *clock* rings *an alarm clock* rings" Own corpus: learner further specifies the type of 'clock' that he or she is talking about.
A-repair: ambiguous-reference repair	The speaker repairs the referring expression because of ambiguity.	"*the man* er the Guelda's husband is walking in the street." Own corpus: the learner replaces 'the man' to make clear what man he or she is referring to.
A-repair: coherent-terminology repair	The speaker repairs a term to be more coherent with the terms used so far.	"but *this letter* is er the order er your request is er anyway…" Kormos (1999, p. 382): the learner reported he or she decided to use 'order' because he or she had used it before, and so it was better to use it again instead of 'this letter'.
A-repair: pragmatic appropriacy repair	The speaker repairs part of the message to make it more pragmatically appropriate in a specific situation.	"I *what can I do for you*?" Kormos (1999, p. 382): the learners reported he or she had intended to say "can I help you?" but found the repair to be more appropriate in the situation.
A-repair: repair for good language	The speaker repairs his or her message in order to use more sophisticated language.	"C'est qu'un con, un idiot pardon." Brédart (1991, p. 127), provided by Kormos (1999, p. 318), in which the learners replaces a word for what he or she thinks is a better one.
Error repair: lexical	The detection of a lexical error causes the speaker to self-correct.	"and a man and a other man enter to the to the room who after wa bueno before was Guelda" Own corpus: learner replaces the wrong adverb by correct one.
Error repair: syntactic	Speaker self-corrects after detecting a syntactic error.	"a woman *that she lives* in the no *that lives* in the house". Own corpus: speaker eliminates extra subject.
Error repair: morphological repair	Speaker finds a problem with the morphology of the word and repairs it.	"the man go go went went into the room." Own corpus: learner is narrating a story in the past tense, detects an error, and repairs it.
Error repair: phonological repair	Speaker detects a phonetic error and repairs it.	"the woman er gets into the the wait ([wait]) the wait ([weit]) room." Own corpus: learners detects an error with a diphthong and replaces it by a different one.

1.5 The role of attention and memory in language production

Excellent and thorough reviews have analyzed the different conceptions of attention and memory in general as well as in relation to second language acquisition (Ellis, 2001; Robinson, 1995a; Robinson, 2003b; Schmidt, 2001). Rather than reproduce a summarized version of such extensive reviews, the next two sections will focus on those aspects of attention and memory that are most directly relevant to the relationship between task complexity and L2 performance, which is the central concern of this study.

Language processing can take place for comprehension, production, and learning. Deliberately, the description of attentional and memory processes will be biased towards production in this study. Inevitably, references to comprehension will be made since comprehension, as we saw in Levelt's model, is an integral part of production without which the latter could not occur. Also, although learning is not a necessary consequence of engaging in language comprehension and production, it is often associated with them.

Within the field of SLA, there has been a growing interest in these two constructs since they have been seen as the processes responsible for not only comprehension and production but also interlanguage development. Models of attention and memory have come from cognitive psychology.

1.5.1 Attention

When describing L1 models of language production, it was seen that both procedural and declarative knowledge are "attended to" during message conceptualization. However, the existence of attentional and memory mechanisms underlying production is assumed by such models, and therefore they often go unexplained.

Put in a hierarchical manner, attention is a mechanism which is part of short-term memory, which in turn is the currently activated part of long-term memory. SLA theories have traditionally drawn on cognitive psychology theories that viewed attention as the executive control process that directs the serial passage of information from short-term memory (STM) to long-term memory (LTM) (Broadbent, 1958). These theories have recently been challenged by connectionist accounts of information processing, which see attention as being distributed throughout the entire processing system, and not on just one single executive control. In turn, skill development models like Shiffrin and Shneider's (1977) theory of automaticity or Anderson's (1983) ACT theory of skill acquisition which have inspired much of SLA research, are now being complemented (and sometimes challenged) by models of the study of action such as Wickens' (1989) model of multi-task performance. As will be further detailed in the following sections, attention involves selection, capacity, and effort or sustainability in information processing.

Information processing concerns how information is transformed from the moment it is perceived by a sensory register to the moment it enters short-term memory and, eventually, long-term memory, and how it then leads to response selection and execution. So information goes through the phases of perceptual encoding, central processing, and responding. Since people are exposed to enormous amounts of information, there must be some kind of mechanism that helps us select only part of the incoming information for further processing. The mechanism responsible for information selection is attention.

Most models of attention would agree that there exists a sensory register which first accommodates the incoming input[19], a detection device which acknowledges that the information has come into the system for further processing, and short-term memory, which is responsible for semantically processing the selected information. Attention also supervises response selection and execution However, the problem is precisely with how and when information is selected.

1.5.1.1 Selection

For years, the debate has been centered on the point at which selection takes place. In order to understand how selection works, we can imagine a sound mixing

[19] Note that input is used here in its most general sense, and it is not restricted to linguistic input yet.

board[20] in which we have a number of channels that can be potentially 'on' or 'off' (e.g. 1 through 10) for recording.

From a limited-capacity conception of attention, researchers like Broadbent (1958, 1971) have suggested that selection happens early after information hits the sensory register. A selective filter or 'bottleneck' only allows partial analysis of some specific features of the input to take place, the filtered information is passed on to the detection device, and it is then processed for meaning in short-term memory. In the mixing board image, only one channel will be on so only the sound of one instrument will come to be recorded (See Figure 6 on the following page, model A). In other words, early in the process, information is "filtered" for further processing on the basis of limited attentional resources. Other proposals like the one by Treisman (1964) have proposed the existence of an attenuation filter, which allows for processing of both sensory and semantic information. Following the mixing board simile, two channels can be on, so the sound of two instruments can be recorded, although each channel may be recorded at a different level of volume (See Figure 6 below, model B). For others (Allport, 1987; Neumann, 1996; Sperling, 1960) several sources of information can be processed in parallel, and selection takes place later in working memory, after full semantic processing has taken place. In the image used so far, several channels would be on, and so the sound of several instruments and voices would be recorded (See Figure 6 below, model C). In this

[20] Sound mixing boards are used in recording studios for musical recordings in which several instruments and voices are involved. Normally, one single channel is assigned to each instrument and to each voice, and each channel can be regulated for volume. Whatever sound goes through the channels that are 'on' will be recorded.

view, selection is not the functional consequence of limited attentional capacity but a consequence of action control.

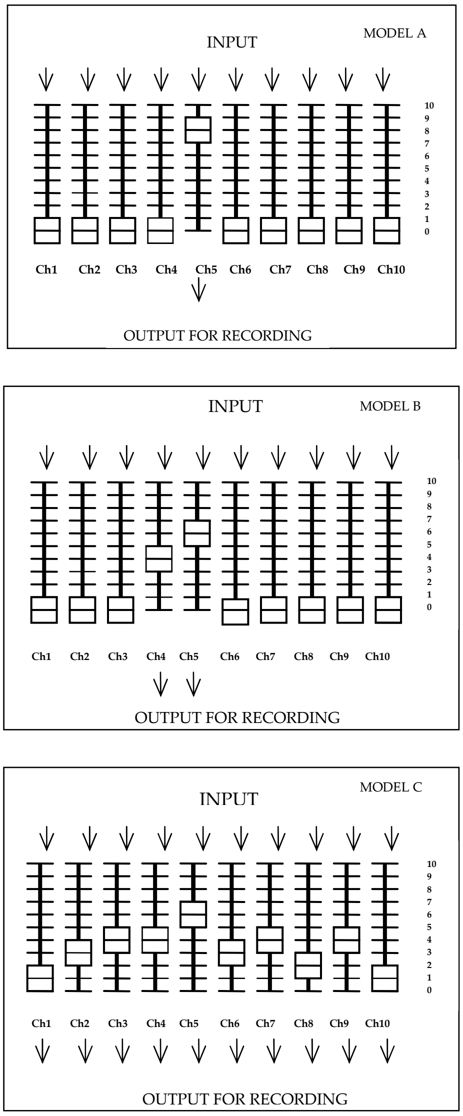

Figure 6. Three models of selection of information ranging from early selection (Model A) to late selection (Model C).

In the case of production, certain bits of information in the repository of knowledge need to be attended to and selected. From the flow of thought, selection of information must take place in order to generate a communicative plan.

1.5.1.2 Noticing

Closely related to the idea of selection as a function of attention is Schmidt's (1990, 1993, 2001) proposed construct of 'noticing'. Although Schmidt's theory of noticing has been questioned by some researchers (Carroll, 1999; Tomlin and Villa, 1994; Truscott, 1998), it stands as a feasible explanation of how attentional mechanisms can account for second language acquisition. The idea behind the construct is that we select a number of features in the surface structure of utterances in the input. Out of all the features we focus on, some are noticed and become intake for learning.

Schmidt (1990, p. 26) defines intake as the part of the input that learners notice:

> "I use noticing to mean registering the simple occurrence of some event, whereas understanding implies recognition of a general principle, rule, or pattern. For example, a second language learner might simply notice that a native speaker used a particular form of address on a particular occasion, or at a deeper level the learner might understand the significance of such a form, realizing that the form used was appropriate because of status differences between speaker or hearer. Noticing is crucially related to the question of what linguistic material is stored in memory...understanding relates to questions concerning how that material is organized into a linguistic system. "

Tomlin and Villa (1994), however, have questioned the idea that awareness is necessary for acquisition. In their account of attention, they distinguish between alertness, orientation, and detection. Alertness can be explained as the general predisposition or readiness for incoming stimuli. Orientation would consist of activation of higher level schema or an action plan according to the expectations about the type of incoming stimuli. Detection means focusing on just a bit of the information which will be further processed in working memory. For them, acquisition operates at the level of detection and not at the level of selection plus noticing, in an account that advocates that learning can take place without awareness. To this debate, Robinson (1995b; 2003b) adds the idea that the focusing and noticing processes are aided by memory mechanisms such as maintenance and rehearsal which, jointly, are responsible for learning.

So depending on the stance one takes, learning may take place with or without awareness, but all researches agree that attention is needed for SLA to take place. As Robinson (2003b, p. 641) states, even if noticing is not necessary, it certainly contributes to learning and retention. Despite well-attested problems in operationalizing exhaustive measures of noticing and awareness (Allport, 1988; Leow, 1997; Robinson, 1995b; Shanks & St. John, 1994), a number of pedagogical proposals have invoked the construct of noticing to defend the proposed pedagogical values of techniques which focus learner attention on form during communicative activity (Doughty, 2001; Long & Robinson, 1998). Some examples are input enhancement (Sharwood-Smith, 1991), processing instruction (Van Patten, 1996), and recasts (Doughty & Williams, 1998; Doughty, 2001), all of which are likely to be beneficial for learning.

1.5.1.3 Capacity

Even more important for us here than selection is the idea of the capacity limitations of attentional resources. In the SLA field, differentials in cognitive load are often explained in terms of capacity limits. When task demands are made higher, it is often suggested that there are not enough attentional resources to attend to both meaning and form, or to different dimensions of the same task. Although researchers in cognitive psychology agree that some tasks are more attention demanding than others or that performing two tasks is more attention demanding that just performing one, there are very different views about exactly how attention during performance works.

Kahneman (1973), who sees attention as a part of working memory, believes in the existence of a single volume of attention whose limits are dependent on the level of arousal. For example, during an attention-demanding task like an exam, the limits of the single volume of attention may be stretched more than during a less attention-demanding task. The metaphor used by Kahneman is that of consumption. The higher the demands, the more resources from the single volume are consumed and the higher the effort that has to be made. This type of model has been criticized because it fails to explain the phenomenon of divided attention during dual or multiple-task performance.

Wickens' (1989) model of dual-task performance proposes breaking the single volume of attentional resources into a series of dichotomical dimensions. The dichotomy affecting processing stages opposes the perceptual/cognitive dimension (i.e. encoding and central processing) to the response dimension. For codes of processing, the spatial is opposed to the verbal. In the case of modality, auditory perception is opposed to visual perception, and responses can be manual or vocal (See Figure 7 below).

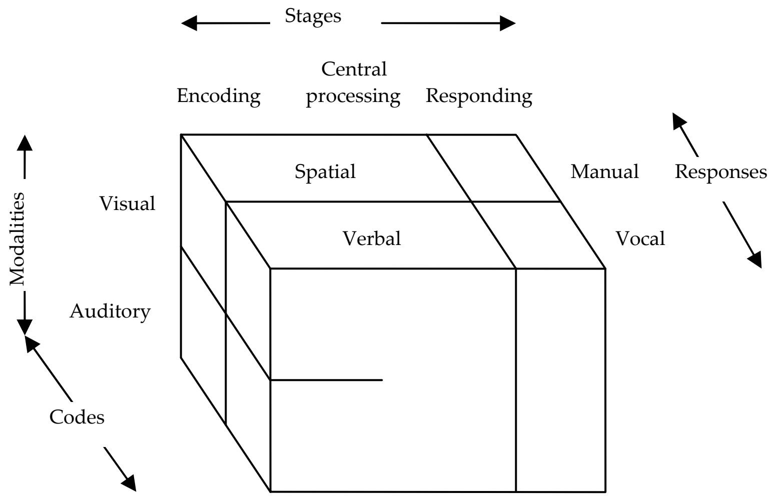

Figure 7. Wickens' (1992) model of multiple resource pools.

Wickens claims that because these different dimensions draw on different resource pools, competition for attention may not necessarily happen. There is competition when two tasks or two dimensions of the same task feed on the same resource pool. For example, having two conversations simultaneously would lead to competition of resources and poor performance since they would both draw on the same ‘verbal’ resource pool. On the other hand, driving home while simultaneously singing or verbalizing thoughts would draw on the ‘manual’ and ‘vocal’ dimensions of performance, and therefore no competition for attention would occur. Thus, the model predicts that the amount of interference will depend on the number of shared levels on all three dimensions (i.e. codes, modalities, and responses). Critics of this model have suggested that Wickens’ resource pools may have been too broadly defined, since evidence exists that two tasks defined as drawing on the verbal dimension, like simultaneous spelling and arithmetic calculation, have been shown not to necessarily enter into competition.

Finally, other models (Navon, 1989; Neumann, 1987) have taken Wickens’ idea of different resource pools of attentional resources but have abandoned the idea of capacity limitations. They have suggested that what leads to less efficient performance is ‘cross-talk’ and interference. Oberauer and Kliegl (2001, p. 187) provide a fine distinction between the two terms. They define interference “as mutual degradation of memory traces that are held in working memory simultaneously”. For example, if two words or two objects are held in memory simultaneously and they share many features, interference may occur. They cite Nairne’s (1990) example of the representation of a new word, for example, that may overwrite the features of an old one if these features are shared among the two. They define ‘cross-talk’ as “the confusion between two elements that are held simultaneously in working memory.” As opposed to interference which affects memory traces themselves, crosstalk may occur “at the selection of one out of several elements in working memory.” In this view, then, attention control is responsible for the decision to engage, disengage, and shift attention between tasks and for the pursuit of intention. When task demands are made higher because the amount of input is increased, because response alternatives are many, or because they are very similar, different codes may compete or there may be ‘cross-talk’ between them. Robinson (2003b, p. 645) provides Shaffer’s (1975) example of typing words that are heard at the same time as typing words that are visually presented. Because the same code is used confusion arises and words are mistakenly typed. A breakdown in attention control results in involuntary shifts of attention. Less effective performance, therefore, is not the result of capacity limitations. Sanders (1998) describes connectionist models of representations, processing, and attention which propose units in a network which are mutually being activated and deactivated. These units work in parallel at different levels of processing, and Sanders (1998, p. 15) states that what these models suggest is that: “processing limits are due to interference, confusion and cross-talk among elements of a neural net and not to capacity constraints”.

The discussion of how attention works during language processing will be specifically dealt with in Chapter III. As we mentioned before, one position (Skehan, 1998; Skehan & Foster, 2001; VanPatten, 1993; VanPatten & Cadierno, 1993; VanPatten, 2000;) implicitly draws on limited-capacity models of attention while another position (Robinson, 1995b, 2001a, 2001b; 2003a; forthcoming) advocates a combined multiple-resource and non-limited capacity interference model of attention and, as will be seen later on, evidence has been provided in both directions.

1.5.2 Memory

This section will focus on two aspects of memory. Firstly, some widely accepted characteristics of memory as well as its architecture will be discussed. In this description, attention will be integrated. Secondly, we will discuss how memory processes have been associated with learning. Of particular interest to us will be the distinction between memory macro-processes and micro-processes, both of which are responsible for development.

1.5.2.1 The architecture of memory

It is a widely accepted idea that different parts of memory serve different functions, and that we can talk about perceptual or sensory memory, short-term or working memory, and long-term episodic or semantic memory. Especially interesting for us here is the role of working memory that Cowan (1993, p. 166) has defined as the: "interface between everything we know and everything we perceive or do". In other words, it is the place where connections are established between incoming information, be it new or already known, and knowledge stored in long-term memory.

As opposed to early models which saw memory as a linear process (Baddeley, 1986; Broadbent, 1984), Cowan (1993, p. 12) has described working memory as the currently activated parts of long-term memory and attentional focus[21]. Robinson (1995b) has advanced a structure which distinguishes between short-term memory and working memory. Working memory is the part of short-term memory which receives focal attention. This implies that the part of the input that has been detected can go in two different directions. On the one hand, without awareness, making use of peripheral attention and, hence, without focal attention, information can briefly enter working memory and automatically activate already stored information in long-term memory. On the other hand, the part of the information that enters working memory and receives focal attention is noticed and rehearsed. Robinson relates maintenance rehearsal, which is data-driven, instance-based processing to implicit learning, while elaborative rehearsal would consist of conceptually-driven, schema-based processing which would be related to explicit learning. Hierarchically presented, working memory is the part of short-term memory that receives focal attention. Working memory and what is left of short-term memory (which only receives peripheral attention) are the activated part of long-term memory (See Figure 8 on the following page).

[21] It should be pointed out that in the case of Baddeley, his position has shifted from a conception of a linear memory process to a position which, like Cowan's, suggests that WM is the currently activated part of LTM.

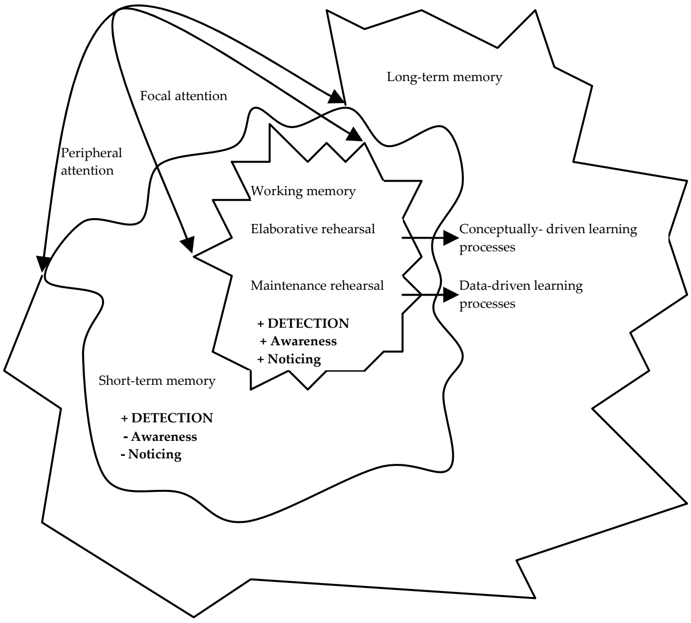

Figure 8. Robinson's (2003b) representation of attention (focal and peripheral), noticing (attention + rehearsal in WM), memory (working memory, short-term, long-term memory) and "detection" (recognition without awareness in passive short-term memory).

In short, language processing starts when input is perceived and automatically processed by the perceptual store of working memory. It then goes into short-term memory, where an automatic process of recognition (detection without awareness or noticing) activates a number of categories which pre-existed to the input. A number of mechanisms compare the similarities and differences between new input and already stored knowledge in long-term memory. In the case of an L2, new input encoding processes have access to interlanguage representations. This, however, is different from learning, since for learning to take place (i.e. for new input to be encoded in long-term memory), no matching of new input with previously existing knowledge must occur. The information must enter focal attention (not just automatic recognition) and enter short-term working memory, and it has to go through two kinds of rehearsal processes, that is, maintenance rehearsal (data driven, instance-based processing) and elaborative rehearsal (conceptually driven, schema-based processing).

Following Doughty (2001), in the next section a distinction between cognitive micro-processes and macro-processes of language processing will be made. Given that micro-processes are dependent on current processing conditions and, therefore, more relevant to our discussion of production, they will be discussed first. The description of more automatic macro-processes, such as internalization of input, form-meaning mapping, analysis, and restructuring will follow.

1.5.2.3 Memory and learning

We have mentioned at different points that language processing can take place for comprehension, production, or learning. Doughty (2001) identifies two micro-processes of memory which can potentially contribute to learning. These are selective attention and cognitive comparison. Selective attention was discussed in Section 1.4.1.1. We saw that Schmidt (2001; Schmidt & Frota, 1986) suggests that for learning to happen, learners need to notice the gap, that is, discover what they need to say but cannot say so that they can look for that knowledge in the input later on. In Doughty's (2001, p. 225) words: "For it to be cognitively possible for learners to notice gaps, they must have sufficient and coordinated working and long-term memory resources to enable the cognitive comparison". That means that the learners need to have enough resources to be able to compare their own utterances, or their failed attempts at building them due to insufficient knowledge, to the utterances of their more competent interlocutors. In other words, learners notice the gap, then at some other point they notice the solution to their gap in their interlocutors' input, and they make a cognitive comparison which restructures their interlanguage. As Doughty (2001, p. 227) adds: "making connections between the known and the unknown eventually leads to knowledge restructuring." In the architectural structure of memory, then, selective attention and cognitive comparison happen within working memory and with awareness, and they have been seen as two processes which are responsible for learning.

As far as macro-processes are concerned, Doughty (2001) describes them as automatic and rather inaccessible. She identifies four macro-processes: internalization, mapping, analysis, and restructuring. Given that these processes are not directly relevant to the goals of this dissertation, they will only be briefly outlined here. A definition of each basic concept is provided and references to their advocates supplied (See also Table 6 for a summarized presentation of both micro- and macro-processes).

In order to understand the first construct, the internalization of input, we can resort to two key concepts in the input literature: 'comprehensible input' and 'intake'. Regarding the first concept, in his *i* + 1 hypothesis Krashen (1982, 1985) claimed that the development of learners' interlanguage is the result of comprehending input that contains linguistic features one step beyond their current level of competence. In this model, the use of context and extra-linguistic information leads to the comprehension of new linguistic material. In his view, exposure to comprehensible input alone leads to acquisition. Regarding the second concept, several researchers (Corder, 1967; Gass & Selinker, 1994; VanPatten, 1993; 2002; VanPatten & Cadierno, 1993) have advanced and defined the concept of 'intake'. Gass and Selinker (1994, p. 303) have defined intake as the process of assimilating linguistic material, by which new input is matched against prior knowledge, and processing occurs against existing internalized grammar rules. Van Patten (2002, p. 757) has defined it as "the linguistic data actually processed from the input and held in working memory for further processing". Abundant research has made use of such conceptualization, and VanPatten's (1993, 2002) model of 'processing instruction' stands as an example of how the internalization of input can be fostered.

Secondly, the concept of mapping originated in studies of first language acquisition and has been incorporated into the explanation of second language acquisition. L1 acquisition models have hypothesized that what leads to language acquisition is the mapping or linking of basic cognitive notions to a fixed set of phonological forms which children extract from the input they receive. Following Slobin (1985), Doughty (2001, p. 218) argues that: "what ensures the mapping of forms to meaning and grammatical functions is the support of the physical context" in a process that assumes learners' predisposition to systematization.

Thirdly, and related to the construct of mapping, the concept of analysis has been presented as an input-independent construct. Bialystok (1994, p. 561) defines analysis as "the process by which linguistic and conceptual representations become more explicit, more structured, and more accessible to inspection. Analysis proceeds on implicit unstructured representations and converts them into an increasingly explicit form…it is the means by which cognitive processes are responsible for altering mental concepts of grammar in the absence of any ongoing input or correction". Both mapping and analysis are seen as continual and cumulative processes.

Finally, restructuring is seen as an abrupt change by means of which some specific aspect of interlanguage becomes more efficient in mental representation. DeKeyser (2001), for example, has used evidence from learning curves and reaction times of explicit instruction to show that restructuring can take place after extended practice. DeKeyser has argued that declarative knowledge becomes procedural knowledge, and, in a second step, automatization of the rule occurs. For Doughty (2001, p. 22): "Proceduralization is a process that results in a more efficient production rule."

These four macro-processes of internalization of input, mapping, analysis, and restructuring are seen as responsible for language acquisition (See Table 6 on the following page). With the definition of these constructs, an attempt has been made to give an answer to the third question we posed at the beginning of this chapter regarding how attention and memory mediate production and development, and it is to the answering of the last question that we now turn.

Table 6

Micro and macro memory processes leading to second language learning and acquisition (based on Doughty, 2001).

Processes	Constructs	Definition	References
Micro	**Selective attention / focus / noticing**	a shift of attention from meaning to form processing	Schmidt (2001) Long and Robinson (1998)
	Cognitive comparison	in processing for language learning learners monitor their intentions, input, and output	Doughty (2001)
Macro	**Internalization of input / intake**	a process by which some data in the linguistic input are held in WM for further processing against prior knowledge (e.g. grammar rules)	Corder (1967) Gass and Selinker (1994) Van Patten (1993, 2002) Krashen (1982, 1985)
	Mapping	the continual and cumulative process of linking concepts to phonological forms and grammatical functions	Slobin (1985)
	Analysis	the continual and cumulative process by which linguistic and conceptual representations become susceptible to inspection because they become more explicit, structured, and accessible	Bialystok (1994)
	Restructuring	the abrupt process by means of which some aspect of interlanguage becomes more efficiently represented in the learner's mind	DeKeyser (2001) Anderson, Finchman and Douglass (1997)

1.6 Language production and interlanguage development

So far we have seen some crucial processing mechanisms involved in language production. We have described the different components involved in L1 production as suggested by Levelt (1989), followed by the processes of production which are specific to speaking a second language. As underlying production, the processes of attention and memory have been outlined and associated with learning and acquisition. We have also seen that the processes involved in comprehension, production, and processing for learning are closely interconnected. Hence, we have answered the first and second questions we posed at the beginning of this chapter.

Nevertheless, we have not answered the last question, that is, we have not discussed any proposal coming from the SLA field that specifically connects production and interlanguage development. This will be the aim of the subsequent sections, which will draw on the work of Swain (1985, 1993, 1995, 1998), Skehan (1998, Skehan & Foster, 2001), and Robinson (2001a, 2001b, 2003a, forthcoming) to outline the potential connection between production and interlanguage development.

1.6.1 The need for output

The work of Swain (1985, 1993, 1995, 1998) in the last two decades has raised the interest of researchers in output as a potential cause of L2 development. In part, Swain's interest in production was a reaction to models of language learning that postulated that exposure to input was sufficient for language learning. Researchers like Krashen (1985, 1989) had argued that output is in fact the sign that a progress in language acquisition has taken place, and that production does not contribute to second language acquisition in any way. Canadian immersion programs (Swain, 1998, p. 65-66), in which learners are exposed to rich and varied input, show evidence that despite successful advances in comprehension skills and communicative fluency, learners tend to be grammatically inaccurate. Being exposed to the input is essential for language learning but not sufficient to acquire an almost native level in the L2 despite favorable conditions. As Swain found out, learners can get through the system with incomplete knowledge (i.e. grammar inaccuracy) because in that kind of system, Canadian immersion programs, they are rarely 'pushed' to be more accurate, and because teachers do most of the talking in class.

In such a context, Swain (1985) advanced the concept of comprehensible output. The rationale behind it is that when in interaction learners are faced with communication problems, they will be 'forced' or 'pushed' into being more accurate and precise with their language, and more coherent or clearer in their messages. By experiencing problems with communication, that is, when learners use the language and find difficulties in being understood, they must pay attention to their own language in order to communicate their message in a successful way. These ideas gave rise to Swain (1993, 1995) "Output Hypothesis", which has proposed a series of functions for output that potentially lead to interlanguage development.

1.6.2 Swain's Output Hypothesis

According to Swain, the first function of production is hypothesis-testing. As learners use the language, both individually (e.g. a class presentation in which immediate feedback is not available) and in interaction with others (i.e. when feedback can exist), they test their own hypotheses against their receptive knowledge. In this sense, De Bot (1996, p. 551) suggests that when learners speak the L2 they can resort to receptive knowledge, "which is assumed to be more stable and reliable than productive knowledge", which can lead to a better, more accurate, and appropriate performance. Learners may extend L1 meaning to L2 contexts, they may extend L2 meanings to new contexts, and can engage in IL system hypothesis formation and testing (Swain & Lapkin, 1995). In interaction, they can test whether they are understood and whether their interlanguage is linguistically well-formed. By receiving feedback from their interlocutors, learners engage in negotiation of meaning, and their production is pushed to produce more accurate and precise language. If we accept the argument that students only learn what they are ready to learn at the right developmental stage (Pienemann, 1985), we must also accept that the input that students receive at a certain time may or may not be good for language development. Producing language, on the other hand, allows the students to monitor their own learning process. From an interactionist perspective, Gass (1997) subscribes Swain's suggestion of output as a generator of hypotheses, and the idea of output as a generator of feedback that can be used to accept or reject such hypotheses, and adds the ideas that output helps to develop automaticity of interlanguage production, and it also forces a change from meaning-focused processing of the second language to the syntactic processing mode. Regarding the two latter functions of output, automatization and syntactic processing mode, she cites McLaughlin (1987, p. 134) who defends the idea that automatization "involves a learned response that has been built up through the consistent mapping of the same input to the same pattern of activation over many trials".

Another function that Swain suggests is the metalinguistic function of engaging in production. 'Metalinguistic' should not be understood as using linguistic terminology to think about language. They idea is that as they talk, learners can "control and internalize linguistic knowledge" (Swain, 1995, p. 126). According to Swain, learners engage in syntactic processing that goes beyond the need to be understood. In a genuine communicative context, thinking about rules, forms, and form-function relationships may lead learners to produce modified or reprocessed output which in turn may contribute to language acquisition. In Swain's view (1998), encouraging students to use metatalk when faced with a linguistic problem should help them learn about language learning processes, because it shows "language in progress". As Swain herself has shown (1998), however, when students get involved in metatalk, because the task itself demands it, they tend to learn both correct and incorrect solutions to language-related episodes.

A third function has been identified as the "noticing/triggering" function of output. This function is indirectly related to input. Following Schmidt's concept of "noticing the gap", Swain suggests that students may notice the gaps between what they want to say and what they can actually produce, therefore realizing the knowledge they lack about the target language. This recognition of gaps in knowledge can be internal, so that it is the student himself or herself that notices the gap, or external, that is, detected by the interlocutor or teacher. The noticing function of output may have two subsequent consequences. If input is not immediately available, learners may engage in a series of thought processes by which they revise their knowledge to either consolidate it or to generate new knowledge. Secondly, if input is immediately available, students may pay more attention to subsequent input in order to solve their problem and fill their gap. In her paper of 1998, Swain showed that noticing the gap actually happens as a consequence of engaging in output. She calculated the number of language-related episodes (e.g. the operationalization of metatalk) that provided evidence that students pay attention to what they need. This function, however, has been tested by Izumi et al. (1999) without satisfactory results. They showed that there is not necessarily a direct link between noticing and immediate incorporation and subsequent language learning. More specifically, what Izumi et al.'s work suggests is that if a link exists between noticing during production and development, it is an indirect one mediated by other factors such as task complexity or effective focus on form (i.e. more efficient comparison of interlanguage output and target language input). The noticing function of output, then, has been proved difficult to investigate, and only inconclusive evidence has been supplied so far.

1.5.3 Skehan's additional functions of output

Skehan (1998) has identified three more functions of output which can be added to the hypothesis-testing, metalinguistic, and noticing functions advanced by Swain. Skehan's first proposal presents output as a generator of better input. The assumption is that being exposed to input from the interlocutor does not guarantee that that input will be high quality. Thus, in interaction, when the listener provides feedback that the input is not being understood, the speaker adapts by reformulating the input in a way that will be more adequate to the listener's current level of competence. In that way, feedback generates better quality output[22]. An advocate of this position which has been regarded as 'negotiation of meaning' is Pica (1994), who suggests that clarification requests, confirmation and comprehension checks eventually lead to interlanguage development. This idea is, nonetheless, quite controversial. On the one hand, Skehan (1998, p. 20), admits to the fact that negotiation episodes may encourage hypothesis-testing and syntactic processing. He says, however, that instances of negotiation of meaning may in fact distract students from form, because their cognitive processing system is overloaded by their trying to solve a communication problem, and there may not be

[22] Let us point out here that Skehan's picture is not quite complete, since other problem-solving mechanisms may be applied, such as abandoning the message or avoiding problematic areas, as we saw in Sections 1.4.3.1, 1.4.3.2, and 1.4.3.3).

sufficient time for changes or improvements in interlanguage to consolidate. He concludes that, in fact, for most students, 'negotiation of meaning' does not have as a consequence interlanguage change and development.

On the other hand, Mackey (1999) has provided evidence that engaging in conversational interaction can lead to second language development. More specifically, her work has shown how interaction can contribute to the production of developmentally more advanced structures, hence contradicting Skehan's claim and lending support to the Interaction Hypothesis (Gass, 1997; Long, 1981, 1983, 1985).

In addition to generating better output, engaging in production may have an automaticity function. This function has to do with the idea of skilled behavior. The more we practice speaking, the less of an effort we have to make when we engage again in interaction (DeKeyser, 2001). Psycholinguistic models of learning have shown that the more often we connect the components of utterances, the less of an effort will have to be made the next time we have to produce them. By automatizing speech, we will not need to pay so much attention to how we are communicating a message (i.e. formulation) and will therefore be able to concentrate on what is being communicated (i.e. conceptualization). Skehan (1998, p. 18) suggests, however, that this may apply more to some areas of language than to others (e.g. more to morphology and less to word order). Skehan (1998, p. 21) is again cautious about the contribution of developing strategic competence to interlanguage change and development.

Finally, Skehan suggests a discourse skills development function. Not only by participating in discourse can a learner become a better communicator. Linguistic anthropology (Duranti, 1997) has shown that to become a competent speaker in a community, one must master not only the code but also the system or rules underlying conversation. Connected to this, there is the development of one's own voice within the speech community.

As opposed to Swain, then, Skehan (1998, p. 27) is in fact quite skeptical about the contribution of speaking to interlanguage development, and from a limited-capacity conception of attention, he states that: "The central point is that language use, in itself, does not lead to the development of an analytic knowledge system since meaning distracts attention from form". As we will see, this assumption is questioned by other proposals[23].

1.5.4 Robinson's Cognition Hypothesis and interlanguage development

In his Cognition Hypothesis of adult task-based language learning, Robinson (2001a, 2001b, 2003a, forthcoming) has advanced a series of assumptions and predictions of how increasing the cognitive complexity of language learning tasks may lead to interlanguage development. Both the concept of Task Complexity and the Cognition Hypothesis will be further discussed in the next chapter. However,

[23] The work of Schinichi Izumi and colleagues (Izumi et al. 1999; Izumi, 2002) has revolved around the issue of how beneficial it is to engage in language production. They have addressed the issues of how output contributes to improved communication, how it contributes to interlanguage development, and how it relates to input when learning is concerned, in other words, how input and output interact and how that interaction contributes to learning.

there are two related predictions of the Cognition Hypothesis that are worth outlining at this point, given their relevance to our discussion.

In the first place, Robinson (2003a, p. 52) proposes that increasing the cognitive demands of language learning tasks can recapitulate or reproduce "the ontogenic course of conceptual development in childhood". As Slobin (1985) has suggested, in L1 acquisition cognitive and conceptual development are the driving force behind linguistic development. Robinson believes that this applies to L2 acquisition as long as we accept the fundamental difference between children and adults, which is that in the case of the latter "all the cognitive machinery is in place", whereas it is not in the case of children. In Robinson's view, then, increasing cognitive complexity along resource-directing dimensions (See Sections 2.4.2 and 2.4.3) such as the number of elements in a task (e.g. from the description of a space with few topological marks and relations to a space with many topological marks and relations among them), the degree of displaced past time reference (e.g. tasks in the Here-and-Now or the There-and-Then), or the reasoning demands (e.g. from narrative description of tasks with successive actions that do not require causal reasoning to those which do) leads to second language development.

Secondly, Robinson (2003a, p. 52) assumes Slobin's (1985) idea that adults retain a scale of conceptual complexity acquired in childhood, and predicts that increasing the cognitive demands of tasks may provide learners "with optimal, ontogenetically natural, contexts for making form-function mappings necessary to

L2 development". These parallels in adult-child language development can help operationalize sequencing decisions.

This review of Swain's, Skehan's, and Robinson's proposals, then, has been an attempt to provide an answer to the last question proposed at the beginning of this chapter, that is, how language production can potentially lead to second language development.

1.6 Summary of Chapter I

Chapter I has laid out the basic constructs necessary to understand how language production works and its potential effects on learning and acquisition. It began by outlining Levelt's model of L1 production, which was presented as a model divided into five main components. Message conceptualization, formulation, and monitoring received special consideration. This was followed by the examination of the models that account for the differences between L1 and L2 production, and again the generation of concepts, their formulation, and monitoring were emphasized. Underlying those models and crucial to accounts of second language acquisition, the processes of attention and memory were outlined. As far as attention was concerned, selection and capacity were presented as concepts key to understanding explanations of how the different dimensions of language performance may interact. The structure of memory was outlined and the memory mechanisms responsible for interlanguage development were briefly discussed. In the last part of the chapter, a number of speculations about how production can contribute to learning were summarized.

One of the key concepts we can draw from this first chapter is that language processing happens for comprehension, production, and learning. In the case of learning a second language, these three processes interact in ways that are still little understood. In Chapter I, then, language production has been presented in general, without any connection to how it occurs in instructional settings, which will be the subject of Chapter II.