# Are simultaneous interpreters subject to the central processing bottleneck during language production?

Longjiao Sui

Bachelor of Arts
Master of Arts

A thesis submitted in fulfilment of the
requirements for the degree of
Master of Philosophy

Department of Linguistics
Faculty of Human Sciences
Macquarie University
Sydney, Australia
October 2016

# Acknowledgements

This thesis owes many acknowledgements and gratefulness to many people for its completion. It is their generous and helpful support that allowed me to finish this study within a year and meet my scholarship deadline.

I would like to express my sincere gratefulness to my principal supervisor Dr Haidee Kruger. She is really fantastic and was willing to support me to conduct research that I am actually interested in, and continuously offered her unwavering support for me to complete it. I do believe it is worth my respect and applause for a supervisor who is patient and persistently offers guidance, especially in the face of occasional differences of opinion as a result of different discipline backgrounds. This project could definitely not have been completed without her generous support. In addition, I would also like to express my thanks to my two associate supervisors, Associate Professor Jan-Louis Kruger and Dr Helen Slatyer. They provided much-needed help in participant recruitment and occasional feedback on this thesis.

I would like to express my appreciation to Professor Zenzi M Griffin from the University of Texas at Austin, and Professor Victor Ferreira from the University of California, San Diego for generously offering their testing materials when I was desperate for parts of the unpublished work. I have to say their emails brightened those gloomy mornings and helped me strive harder in preparing and conducting the experiments.

I would also like to thank all the professional simultaneous interpreters for being willing to participate in this study and offering their kind help to pass on my advertisement. It is delightful to get a chance to meet those experts, observe their responsible performance, and hear about their valuable interpreting experience. The completion of this thesis is also the result of their generous support, since it is notoriously difficult to recruit professional interpreters in studies such as these. I would also like to extend my thanks to all my other bilingual and monolingual participants who agreed to participate in a study that does not have direct bearing on their lives.

Finally, I would like to thank my dear friends, Dr Yanan Sun, and Yu Li (PhD candidate). The conversations and discussions with them broadened my knowledge on related research fields and sharpened my thoughts. Furthermore, the valuable suggestions from Dr Yanan Sun on programming is one of the important reasons inspiring me to master a new program in one short month. I feel really fortunate to have met these friends and wish them all the best in their lives and their research.

# Abstract

Simultaneous interpreting (SI), the mode of interpreting generally used at international conferences, is an exceptionally complex language-processing task, which requires interpreters to continuously receive source language (SL) utterances, store the information in memory, transcode utterances into the target language (TL), and articulate the previous segments. What makes SI particularly difficult to perform is that all the above processes are carried out in real time.

Almost all models of cognitive processing during SI share the assumption that cognitive resources are limited and shared by all the components of processing that need to occur concurrently (Gerver, 1976; Gile, 1997; Liu et al., 2004; Christoffels & De Groot, 2004; Padilla, Bajo & Macizo, 2005). This assumption has gone largely unchallenged, and yet psycholinguistic research on language performance has suggested that language production is subject to a central processing bottleneck (Welford, 1952) which allows only one task to be performed at a time.

This raises the as-yet-unanswered question of whether interpreters, despite their professional experience, are also subject to the central processing bottleneck during language production, or whether interpreters' extensive experience leads to more efficient lemma and phonological word-form selection than is the case for bilinguals and monolinguals, alleviating the bottleneck. In this study, I investigate these questions using an experiment in the dual-task paradigm. The results suggest that even language experts such as simultaneous interpreters are subject to the central processing bottleneck during word production. No significant difference was found between the three matched groups of professional interpreters, bilinguals and monolinguals on the duration of the bottleneck stage during word production. In addition, the results indicate that interpreters are as good at anticipating the upcoming word as monolinguals, and better than proficient bilinguals.



**Keywords:** simultaneous interpreters, bilinguals, monolinguals, lexical access, Psychological Refractory Period, dual task.

## Certificate of Originality

I hereby declare that this work is the result of my own research and that the experiment has not been previously submitted for any other degree. The study reported in this thesis was conducted by myself, except for the use of the materials which have been acknowledged.

Longjiao Sui

# Table of Contents

# List of Figures

# List of Tables

# List of Abbreviations

CI: Consecutive interpreting

PRP: Psychological Refractory Period

SI: Simultaneous interpreting

SL: Source language

SOA: Stimulus-onset asynchrony

TL: Target language

WM: Working memory

# Chapter 1. Introduction

## 1.1 Introduction and background

Simultaneous interpreting (SI), the main mode in conference interpreting, is generally acknowledged as one of the most complex language-processing tasks in which humans can engage. SI requires interpreters to continuously receive the utterances from the source language (SL), store its information in the working memory, transcode segments, and articulate the previous utterances into the target language (TL), as well as monitor and correct production errors while listening to subsequent utterances. Although all of these processing tasks are conducted concurrently, each process involves a different information segment, all processed at one time point. Furthermore, interpreters are expected to deliver the interpretation output at a pleasant pace, with few unnatural non-juncture pauses (Cenoz, 1998), with moderate voice (the voice that interpreters produce should not be too high or too low), and under time-induced and circumstance-induced pressure (Gile, 2009). In addition, the content of the interpreting should, and normally does, reflect the propositional content without (much) distortion and omission (Gile, 2009).

The mystery of how interpreters can conduct such complex language processing successfully in real time has been explored by researchers in linguistics (Gile, 2009), psycholinguistics (Gerver, 1976) and neurolinguistics (Hervais-Adelman, Moser-Mercer, & Golestani, 2011). It is important to fully understand how each SI process works individually, as well as in combination. Such understanding is of both theoretical and practical use: it has practical use in developing effective training programmes for SI, and is theoretically important in terms of the further testing and refinement of models of comprehension, memory, language production, and bilingualism in which SI has been investigated (De Groot, 2000).

Lederer, in her analysis of processing requirements in SI, identifies eight "mental operations" that are conducted currently and successively: listening, language comprehension, conceptualisation (i.e. constructing a cognitive memory by integrating

linguistic input with prior knowledge), and expression from cognitive memory (Lederer, 1981, cited in Pöchhacker, 2016, p. 90). The requirement for interpreters to multi-task in order to conduct perception, comprehension, storage, and production (Pöchhacker, 2009, p 55), is broadly reflected in almost all SI models, suggesting that cognitive resources are limited and shared by all the concurrent processing components (Gile, 2009; Christoffels & De Groot, 2005). Professional simultaneous interpreters are supposedly good at effectively dividing their limited cognitive resources and allocating them to each task. The individual processes that are active in SI, such as comprehension (Yudes, Macizo, Morales & Bajo, 2013) and working memory (WM) (Köpke & Nespoulous, 2006), have been explored widely by comparing the performance of professional interpreters, proficient bilinguals, and/or novice interpreting students, and occasionally even with monolinguals. Professional interpreters have been shown to outperform other groups in most cases (for review, see Liu, 2008).

Interestingly, however, the output performance of simultaneous interpreters has received comparatively limited attention. Although the difficult nature of language production under interpreting conditions has been established, professional interpreters, proficient bilinguals and interpreting students are often found to perform similarly in research results on tasks such as word retrieval (Christoffels et al., 2006). This raises the question of whether **language production** should be considered as a less important factor for SI, as has been proposed. More importantly, the question arises whether there is any difference in language production between professional interpreters, bilinguals and monolinguals, beyond differences in speech errors and response latency.

Language production in daily conversation is already effortful. All people sometimes mispronounce a word or part of the sound, and sometimes people may even suddenly be unable to produce a word. When people pay more attention to these speech errors, they may find that the mis-produced word is usually semantically related to the one they intended to say, and they can tell the meaning of the word when they encounter tip-of-the-tongue situations. All of these common phenomena reflect the nature of word production: multiple stages are involved, and groups of semantically related words are also activated alongside the intended word. Thus, in order to correctly produce a word, the selection of the appropriate word and its sound among all the candidates are also required. The situation for someone who can speak more than one language is even more complicated,

since research results show that both or all of their languages are activated during language production (Van Hell & Dijkstra, 2002). That is, compared to monolinguals, bilinguals[1] also need to select the target language in addition to selecting the intended word, and therefore, bilinguals have disadvantages in language production and require more time to produce a word even in their dominant language than monolinguals (Gollan, Montoya, Fennema-Notestine & Morris, 2005).

More interesting results have been found in psycholinguistic research on language production, showing that during word production, the basic unit of sentence production, lemma (word) selection and phonological word-form (whole sound segment) selection are subject to a central processing bottleneck which results in the tasks being performed one at a time (Ferreira & Pashler, 2002). In other words, contrary to the capacity sharing assumption which proposes that all the tasks can be conducted concurrently by sharing the limited cognitive capacity, these results have shown that in order to fulfil the production of a word, another task which is also subject to the central processing bottleneck can actually not begin but has to wait until the word selection has been completed. This well-established central processing bottleneck assumption has been tested and replicated in relation to divided attention with various kinds of non-linguistic tasks (for review, see Pashler, 1994), showing the reaction time to the second task is prolonged when decreasing the interval between two tasks which require speeded responses. The phenomenon of this prolonged reaction time to the second task has become known as the psychological refractory period (PRP).

Although people are subject to the central processing bottleneck during language production, other tasks may also interfere with each other. For example, having a phone conversation while driving could interfere with the driving (Levy, Pashler & Boer, 2006). But what is the impact of the central bottleneck effect on SI, since interpreters only produce one word at a time? Research results from psychology have shown that memory retention and memory recall are impaired and postponed by this same central processing bottleneck (Rohrer & Pashler, 2003). That is, memory recall and memory retention should also be impacted by single word production. Since, as mentioned above, simultaneous

[1] The lexical access of multilinguals is generally assumed to be based on the same model as for bilinguals (De Bot, 2004). In other words, the lexical access models that account for bilinguals, which will be discussed in the remainder of this thesis, are generally taken also to be applicable to multilinguals.

interpreters store the newly received information in the memory for a short while before recalling it and producing it in the target language, they actually continuously produce the interpreting output at the same time as all those processing procedures. Therefore, it is logical to believe that part of the information that the speaker delivered may be missing from the output of interpreters due to impaired memory retention. Moreover, many unnatural pauses may be expected in the output performance of interpreters because of the postponed memory recall. Furthermore, the speed and accuracy of the non-dominant language production should display increased interference because lexical selection is subject to bottlenecks, as has been found in the case of bilingual participants (Declerck & Kormos, 2012). However, contrary to the hypothesised inference, professional interpreters can provide almost all the important information that needs to be delivered with limited unnatural pauses (Wang & Li, 2015).

## 1.2 Aims of the study

Based on the above brief review, the question arises whether simultaneous interpreters, despite their professional experience, are also subject to the central processing bottleneck during language production. If they are, then the question remains whether their extensive SI experience can lead to more efficient lemma and phonological word-form selection compared to bilinguals and monolinguals, and whether this experience then offsets the effect of the central bottleneck. Therefore, this study aims to explore whether professional simultaneous interpreters are also subject to the central processing bottleneck during language production which might impair the SI performance subconsciously. The study also compares the performance of interpreters with that of bilinguals and monolinguals to further explore whether experience in SI helps interpreters ease the burden of the central bottleneck during language production.

Against the background of the brief background in Section 1.1, and the extensive literature review presented in Chapter 2, the following specific research questions inform this study:

1) Are professional simultaneous interpreters also subject to the central processing bottleneck during language production, despite their professional experience?

2) If this is the case, is the stage they are subjected to the bottleneck shorter compared to bilinguals, and possibly similar to the performance of monolinguals when producing a word?
3) Can the suggested good anticipation skills of interpreters lead to more efficient lemma selection in comparison to non-trained bilinguals and monolinguals?

The hypotheses of the study are set out in more detail in Section 3.2, subsequent to the necessary information provided to formulate the hypotheses in Chapter 2.

## 1.3 Overview of method

In order to answer these research questions, this study mimicked Experiment 1 of Ferreira and Pashler (2002), with three groups of carefully matched respondents: professional simultaneous interpreters, untrained bilinguals and monolinguals. In the experiment, subjects conducted a dual task including a picture naming task (in context), and a non-linguistic sound discrimination task. For Task 1 two factors were manipulated to help explore the difference in lemma and phonological word-form selection: sentence constraint and word frequency. Task 2 was a tone discrimination task which included high, medium, and low pitch sounds. Three stimulus-onset asynchronies (or SOAs, which refer to the interval between the onsets of two stimuli) (50, 150, and 900 ms) were included to manipulate the overlapping of the two tasks. The experiment used in this study is described in detail in Chapter 3.

## 1.4 Conclusion and benefits of the study

The question of whether interpreters also encounter the central processing bottleneck is worth exploring. This research will explore whether language production is a less obvious factor which can impair the performance of SI. Furthermore, a comparison of the language production performance and the dual task performance between professional interpreters, untrained bilinguals as well as monolinguals could provide an indication of whether the matter of language production should also be considered in interpreting training, and provide an empirical assessment of whether interpreters are, as suggested in the Effort model (Gile, 2009), efficient in conducting multiple tasks simultaneously. Moreover, the findings may also offer a basis for testing and modifying current bilingual lexical access

models with interpreters' performance. Most importantly, the findings of this study can provide empirical results to question and challenge the well-known assumption that all SI tasks are conducted simultaneously by sharing limited capacity, as suggested in capacity sharing models (Seeber, 2011; Gile, 2009).

# Chapter 2. Literature review

## 2.1 Introduction

To explore whether professional simultaneous interpreters are subject to a central processing bottleneck during language production, this chapter provides a review of the current research in the field of simultaneous interpreting (Section 2.2), focusing primarily on models of interpreting and psycholinguistic research on interpreting. In this section I also provide a rationale for the study by determining whether this question, or related questions, have been explored previously. In Section 2.3 I summarise the language production models of monolinguals and bilinguals and review the current evidence from research. This section introduces key background to language production, and how it is affected by bilingualism. The concept of the central processing bottleneck is introduced in Section 2.4 and the supporting research results are discussed. In addition, the central processing bottleneck model is also compared with the capacity sharing model, in order to further highlight some of the key contentious issues relevant to the research presented in this thesis. This chapter lays the foundation for the identification of specific research questions and hypotheses in Chapter 3.

## 2.2 Simultaneous interpreting

### 2.2.1 Introduction: Basic concepts in simultaneous interpreting

Interpreters are known to have the remarkable ability of mastering and mediating between two or more languages. Not only do they mediate between their two active languages, but they do this in real time. Simultaneous interpreting (SI) is normally carried out by two or three interpreters taking turns in the booth using purpose-designed audio equipment. Interpreters receive information continuously from a speaker in one language (the source language, hereafter SL) through their headphones and provide a clear interpretation into the target language (TL) after a short lag time into a microphone. The TL speech is received by the audience members through their receivers (see Fig. 2.1). In other words, SI

is a concurrent performance that involves auditory perception of the SL, comprehension, memory retention, translation and TL oral production. The concurrent tasks are represented by Gile (1995) in relation to the incoming SL and TL production as follows:

- Segment A: production
- Segment B: retention
- Segment C: listening and analysis.

While the interpreter is rendering the first segment of the speech (Segment A), she is concurrently retaining the second segment (Segment B) of the source speech in her short term memory and listening and analysing the incoming third segment from the speaker (Segment C).

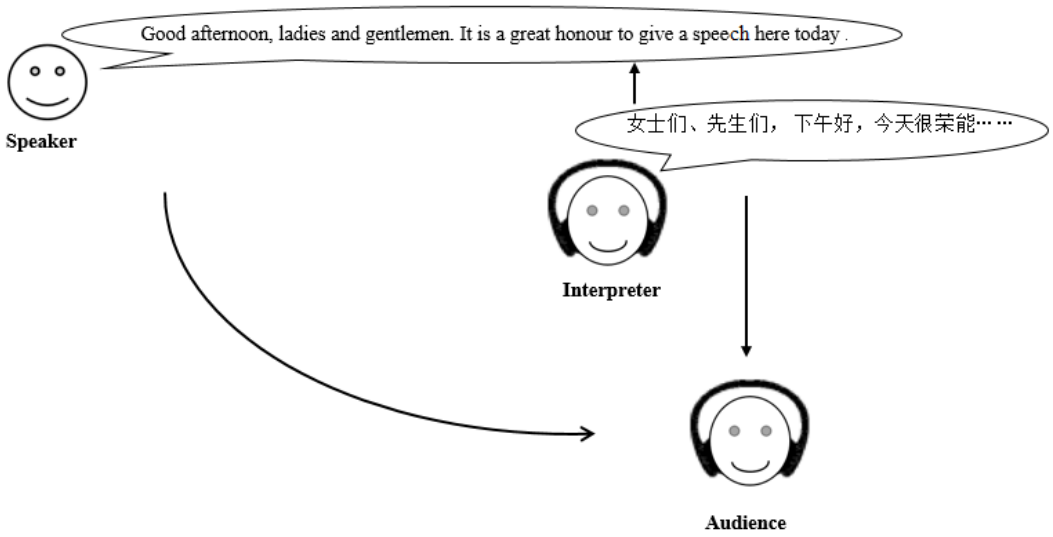



**Figure 2. 1** Simplified representation of SI practice.

It is important to distinguish the different types of interpreting and translation, as De Groot (1997) indicates, because different processing procedures are involved in different tasks, which lead to different performances. Both SI and CI (consecutive interpreting) are modes of interpreting that require oral production, and unlike translation (the written form of language transfer), both SI and CI do not allow control over the rate of information reception and production due to the immediacy of the required response (Pöchhacker, 2016). Consecutive interpreting (CI) allows the interpreter to receive a longer text or around 5 minutes of speech by taking notes, and then to retrieve the information and interpret it into the target language with the assistance of the notes when the speaker has a

pause or stops voluntarily. In other words, in CI the interpreter does not have to deal with language production while receiving and comprehending the new input, and therefore, is not impacted by any potential interference caused by language production, which is a major difference between SI and CI (Gile, 2009).

In contrast, SI relies more on high-speed processing and multi-tasking, while CI relies on more memory retention than SI because the translation unit is longer. Against this background, interesting questions have been raised, such as: How are interpreters able to perform multiple tasks simultaneously? And how do interpreters differ from bilinguals in terms of language processing? To explore these questions, this section is divided into two parts. Section 2.2.2 summarises some prominent models of interpreting that are particularly relevant to this thesis, presenting them in largely chronological order. Section 2.2.2.1 reviews some earlier models, including Seleskovitch's model representing the "Théorie de sens" (Seleskovitch, 1978), Kalina's communicative model of SI (Kalina, 1998), and Setton's (1999) processing model for SI, while Section 2.2.2.2 reviews the basic structure of SI proposed by the influential model of Gile (2009). Section 2.2.2.3 reviews the cognitive load model of Seeber (2011), which is especially pertinent to this study. The current literature on psycholinguistic research on SI is presented in Section 2.2.3.

### 2.2.2 Interpreting models

Several models based on research results or psychological models have been proposed to account for SI (e.g. Kalina, 1998; Seleskovitch, 1984; Setton, 1999), and these interpreting models are briefly reviewed in the following section to provide some general models of the interpreting process.

#### 2.2.2.1 Some earlier interpreting models

**Kalina's model (1998)** is based on a monolingual model of communication. The most salient feature of this model is that production and comprehension in interpreting benefit from word/situation knowledge besides the knowledge delivered by the speaker (see Fig. 2.2). The model indicates interpreting as consisting of the following procedures: The speaker produces the discourse in the SL based on the speaker's mental discourse model. The interpreter receives the information in the SL and transcodes ("translates" in the

model) the message into the TL based on the information. This procedure is also referred to as "communicative mediation", as the intended meaning of the SL speaker must be understood by the interpreter using her own world knowledge. When the interpreter produces the TL rendition for the TL audience it carries the intended meaning of the interpreter. The final receiver of the TL then interprets the meaning according to her own mental model.

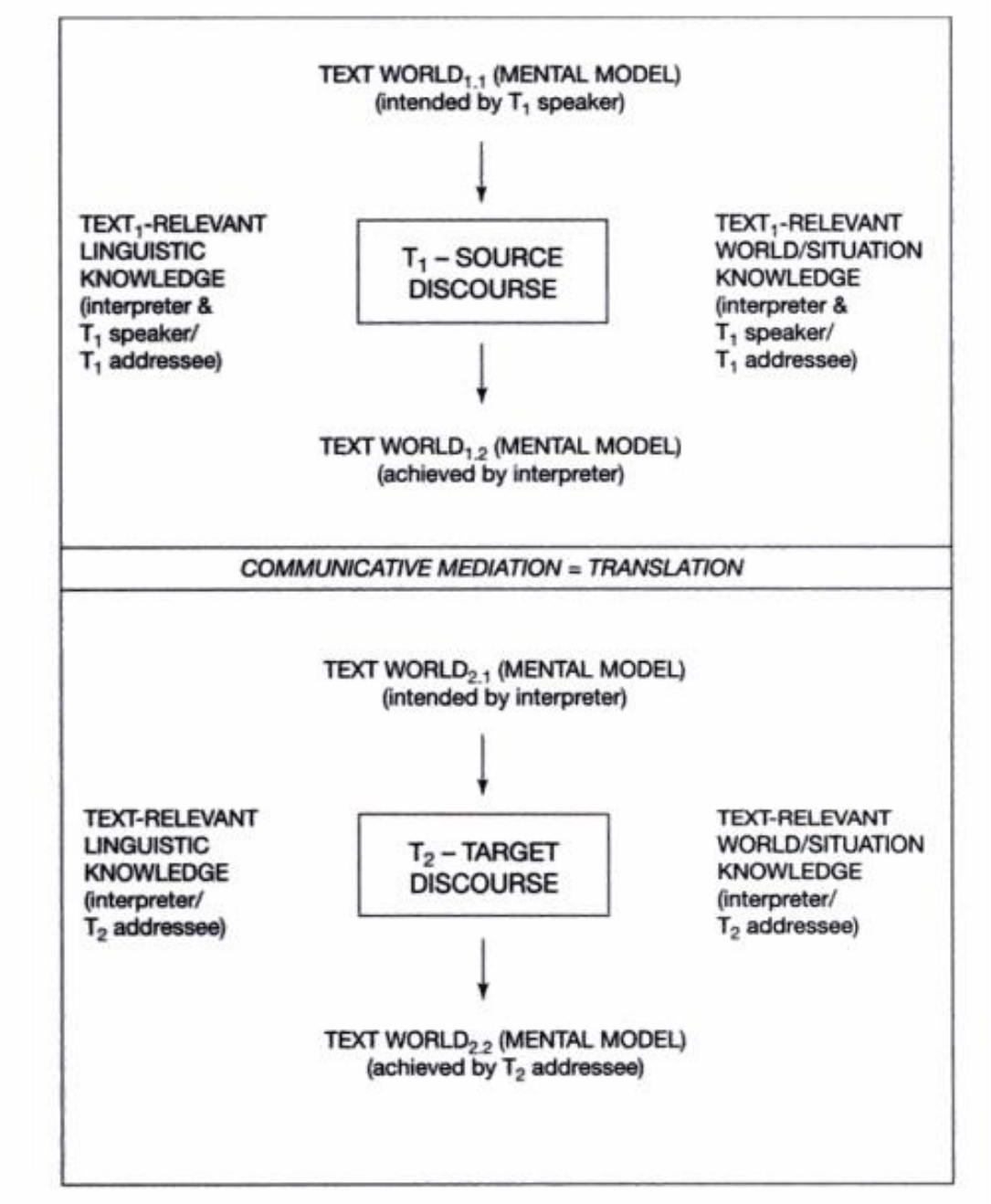



**Figure 2. 2** Kalina's (1998) model of comprehension and production in interpreting. Adopted from Pöchhacker, F. (2016). *Introducing Interpreting Studies*. London: Routledge.

The earliest model of SI comes from the Paris School which proposed the "Théorie de sens" in an attempt to explain the decoding of text into a conceptual "sense" or meaning (Seleskovitch, 1978). In contrast with Kalina's model (1998) which suggests production comes after understanding the information, **Seleskovitch's triangular model (1984)** illustrates that the interpreting mode (relevant to both SI and CI) comes with a different type of "sense", which is suggested as 1) "conscious", 2) "made up of the linguistic meaning aroused by speech sounds and of a cognitive addition to it", and 3) "nonverbal" (Seleskovitch, 1978, cited in Pöchhacker, 2016, p. 89). The "sense" referred to here is correlated with the interpreter's understanding and expression, instead of the literal "transcoding" that was previously considered to be a core element of the process (Pöchhacker, 2016, p. 89). The target language (Language 2) can be activated via both the Language 1 – Sense – Language 2 route, and the direct Language 1 – Language 2 route, where Language 1 can be directly transcoded to Language 2 (see Fig. 2.3). However, if the expression in Language 1 is not "sensible" to the interpreter, then the only route to reach Language 2 production is via the Language 1 – Sense – Language 2 route (see Fig. 2.3).

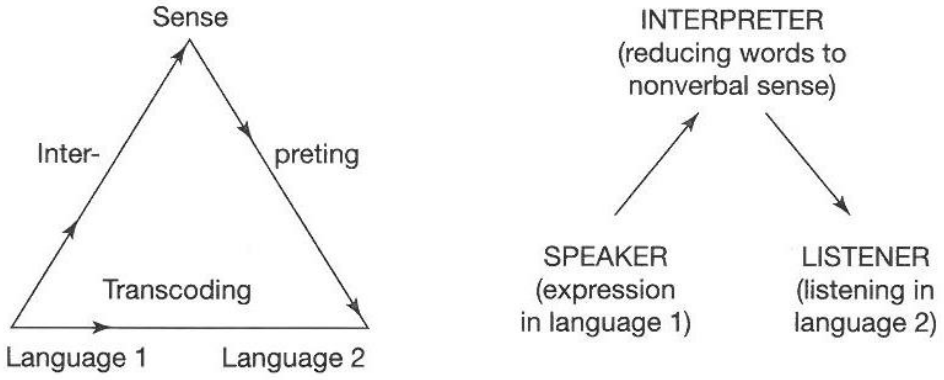



**Figure 2. 3** Two versions of Seleskovitch's (1984) triangular model. The left part of the figure illustrates the Language 1 – Sense – Language 2 route, while the right part represents the Language 1 – Language 2 route. The illustration is adopted from Pöchhacker, F. (2016). *Introducing Interpreting Studies*. London: Routledge.

Compared to Seleskovitch's triangular model (1984), **Setton's processing model for SI (1999)**, which integrates models from cognitive science, provides more detail related to language comprehension, memory, and production (see Fig. 2.4). This model shows that adaptive memory (in the top right of Fig. 2.4) plays a critical role in SI since the situational and world knowledge is suggested "to play an integral part at all stages of cognitive

processing" (Pöchhacker, 2016, p. 96) and it is the intermediation of perception (audio-visual input processing on the left side of Fig. 2.4) to the production (starting with the parser on the right side of Fig. 2.4).

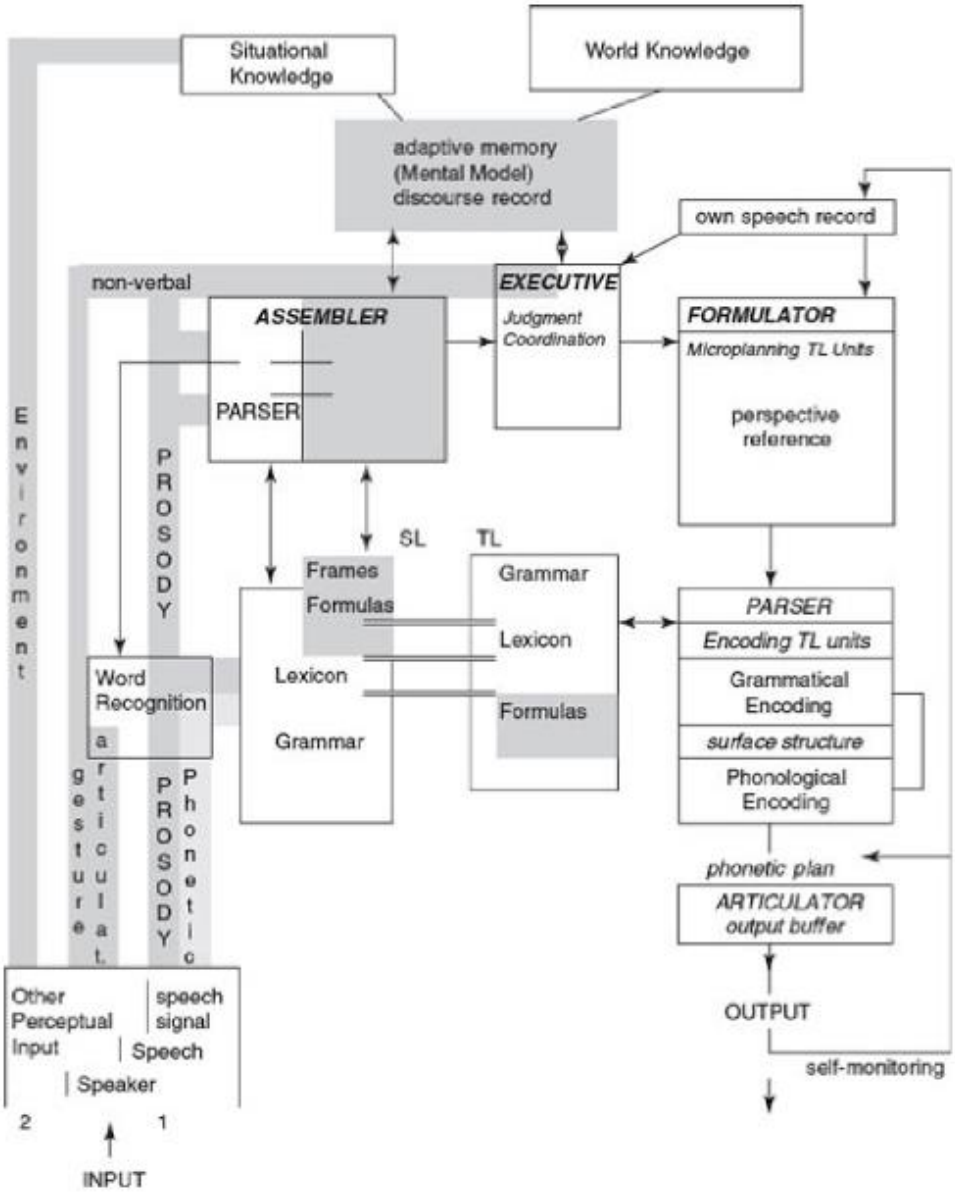



**Figure 2. 4** The processing SI model of Setton (1999). Adopted from Pöchhacker, F. (2016). *Introducing Interpreting Studies*. London: Routledge.

Setton's model closely reflects Levelt's model of speaking but incorporates the SL to TL transfer within the lexicon. Other models based on psycholinguistic theories have been proposed in an attempt to account for SI processing (e.g. Gerver, 1976; Moser, 1978; Darò & Fabbro, 1994). The next step for Interpreting Studies is, as Liu (2008) suggests, to start

focusing on the research results which can offer a ground for testing current opinions on interpreting procedures and abandon incorrect assumptions.

#### 2.2.2.2 The Effort Model (SI) (Gile, 2009)

The most influential and widely accepted cognitive model of SI is Daniel Gile's Effort Model which is, in his own words, "basically a conceptual framework rather than a theory" (Gile 2009, p. 188). This model offers an explanation of the difficulty in conducting SI (there is a separate model for CI which is not discussed here) and outlines the effort allocation for each processing requirement. Before summarising this model, it is important to provide a definition of the word "effort". According to Gile (2009), effort can refer to the mental resource or the decision making that a processing procedure requires: "I called these components "Efforts" to stress their effortful nature, as they include deliberate action which requires decisions and resources." (p. 160). In other words, the processing or the action involved in SI is not fully automatic but actually requires some energy to conduct.

SI, according to the Effort Model, requires a Listening and Analysis Effort (L), which is the effort to receive and analyse the information from the source language; Short-term Memory Effort (M), which is the effort to maintain the analysed information; and Speech Production Effort (P), the effort to offer the interpretation in the target language; as well as an additional effort that works as coordinator for the first three efforts, namely the Coordination Effort (C). Gile represents SI by the formula **SI = L + P + M + C** in accordance with the interpreters' performance of receiving and analysing the new information from the source language (L), memorising it (M), while producing the interpretation of the previous information in the target language (P). Meanwhile, all three efforts are coordinated by the coordination effort (C) (Gile, 2009, p. 168).

For the Listening and Analysis Effort, the effort of receiving the sound wave that carries the speech in the SL, recognising the words, and then analysing the information is surely a non-automatic process that requires effort to conduct. To fulfil the comprehension procedure, not only is the ability of understanding the language itself needed, but additional non-linguistic knowledge, or what is called encyclopaedic knowledge, is also required to assist interpreters to correctly understand the information that the speaker wants to deliver (Kirchhoff, 1976; Chernov, 2004; Schweda-Nicholson, 1987). This assumption

of effort has not only received support from interpreters but is also consistent with research results which show that the accent of the speaker, unclear pronunciation of words, polysemy, dialects, as well as noisy environments may raise the difficulty for interpreters to correctly perceive the information and comprehend it (Gerver, 1969, cited in Gerver, 1975; Gerver, 1974a).

The short-term Memory Effort is the effort required to maintain the segment of information received earlier and analyse it before articulating it. This effort is highly demanding during SI, and its demands can increase when the SL is unclear due to a noisy environment, meaningless information, unclear pronunciation or fast speech (Gile, 2009). This effort can also increase when the TL is difficult to produce because of the difficulty of finding the equivalent term or formulation or suddenly being unable to produce the word. In cases like these interpreters usually decide to continue receiving information while holding the previous item in their short-term memory until they have enough meaningful information to produce or choose an appropriate synonym or formulation. Research results have shown that WM/short term memory has a high correlation with interpreting performance (Christoffels, et al., 2003) and interpreters outperform non-trained bilinguals in free recall WM tasks (Hiltunen, Pääkkönen, Vik & Krause, 2014).

Speech Production Effort is the effort that is involved in finding the right word, and producing logical and clear sentences without (many) grammatical mistakes. A common observation is that language production under interpreting conditions is quite difficult and clearly requires effort. Several factors can contribute to this production difficulty during interpreting. Firstly, there is the difficulty caused by producing the language itself. Almost everyone has encountered the situation of hesitations or pauses, and disfluencies during daily conversation due to suddenly being incapable of finding or producing the appropriate word. This production difficulty not only occurs during spontaneous speech but also during interpreting. Secondly, there is interference from the SL. Simultaneous interpreters continuously receive the SL while producing the interpretation, and the newly arrived SL words may compete with the TL word, and therefore, cause the difficulty of language production (Gile, 2009). Thirdly, there are differing syntactic structures across language pairs and lastly, cultural norms frequently require re-formulation or substitution of the SL by a more appropriate expression in the TL.

These efforts are central to processing in SI, and they are important to consider because the capacity shared by these processing efforts, which can also be allocated voluntarily, is limited. Gile (2009) compares the requirements for each effort with the capacity available. That is, the total capacity required to perform SI can be represented as: **TR = LR + MR + PR + CR** in accordance with the total required processing capacity to conduct SI. (TR) equals the sum of the required processing capacity for listening (LR), memorising (MR), production (PR) and coordination (CR) (Gile, 2009, p. 169). To successfully carry out SI, the **required** capacity for each process should not exceed the **available** capacity for that process. That is, **LR ⩽ LA; MR ⩽ MA; PR ⩽ PA; CR ⩽ CA**, suggesting the capacity required by listening, memorising, producing and coordinating should not exceed the available capacity for these processes (Gile, 2009, p. 170).

If the total demand for capacity exceeds the available capacity, as the formula shows, **TR ⩽ TA** (total available capacity), then interference occurs and performance will be impaired (Gile, 2009, p. 170). However, even if the capacity is adequate for the processing requirements, interference might also occur due to inappropriate capacity allocation. For example, if the interpreter allocates extra capacity to (an) effort(s) (such as monitoring production or allocating more effort to understanding a fast speaker), the remaining capacity is then inadequate to conduct the rest of the task without impacting the performance. Results from experiments have shown that interpreters make new errors and omissions in segments which they previously interpreted correctly when they interpret the same text again, suggesting the difference between the two interpretations is due to differences in capacity allocation (Gile, 1999).

The Effort model provides a valuable model of the experience that professional simultaneous interpreters have during SI. However, contrary to this single-resource model, which suggests that all the concurrent tasks share the capacity from a single limited capacity pool, another influential SI model, the Cognitive Load model (Seeber, 2011), which favours the hypothesis of multiple resources, has been proposed. In the next section, the Cognitive Load model of Seeber (2011) is reviewed in more detail.

### 2.2.2.3 The Cognitive Load Model (CLM) (Seeber, 2011)

The intention of Seeber (2011) in proposing this model was to account for the cognitive load that is generated during SI. In contrast with Gile's Effort model, this model favours the assumption that there are multiple resources and different types of tasks share these different resources. In this case, the empirical result of "perfect time-sharing" (Schumacher, Seymour, Glass, Fencsik, Lauber, Kieras & Meyer, 2001; but see Levy & Pashler, 2001) can be explained. The few basic principles that the model is based on are: 1) more capacity is required to process a dual-/multi-task than to process the tasks involved individually; 2) more interference can be found when the concurrent tasks share the same structure (for more details, see Seeber, 2011, p. 188). For example, Seeber (2011) claims that substantial interference will be found for the combination of language production and perception, since they require the same resources.

Before providing further details of this model, the vectors in which the tasks can be classified need to be outlined. The auditory perception of both input and output is represented by **P**; the verbal-cognitive process in both input and output is represented by **C**; the auditory production of both input and output is represented by **R**; the interference which is generated by combining tasks concurrently is represented by **I**; while the cognitive load which is generated by memory storage is represented by **S**. The total cognitive load generated during SI is represented by adding up these vectors, and can be presented as the formula **Total cognitive load = P + C + R + I + S**. According to this model, the amount of cognitive load remains the same regardless of whether the sentence structure of the TL is the same as that of the SL or different from it. However, the model has some limitations. In particular, it cannot provide an appropriate explanation for the factors that are relevant to lexical access, such as word frequency and cognates.

The CLM considers four situations which often occur during SI, outlining detailed speculations about the amount of cognitive load involved in each situation. When the strategy of *waiting* (in other words, halting production), is adopted by interpreters during SI, the cognitive load will either be alleviated temporarily, because **R** (auditory production) is not involved, or there will be overload if the interpreter continuously receives information from the SL, such as waiting for information located at the end of the sentence, because the capacity required by the rest of the processes (e.g. process **P** and **S**)

increases with SL information input (see Fig. 2.5). When interpreters adopt the *stalling* strategy and pad the production using a slower speed, even though the lag between the SL and TL increases as is the case for the *waiting* strategy, the cognitive load, however, increases since auditory production (R) is involved and overlaps with other vectors (see Fig. 2.5). When interpreters split the interpretation into segments after a rapid consideration and adopt the *chunking* strategy, the cognitive load might increase because of the complexity and difficulty of restoring the original meaning of the SL (see Fig. 2.5). The *anticipation* strategy is mostly recommended by Seeber (2011) since if interpreters can predict the upcoming information from the speaker, they can maintain the cognitive load to the baseline value even if they have to handle the extra interference caused by anticipation. Furthermore, anticipation will also allow interpreters to have a similar lag to the baseline value, since waiting to the end of a sentence is avoided (see Fig. 2.5).

**A**

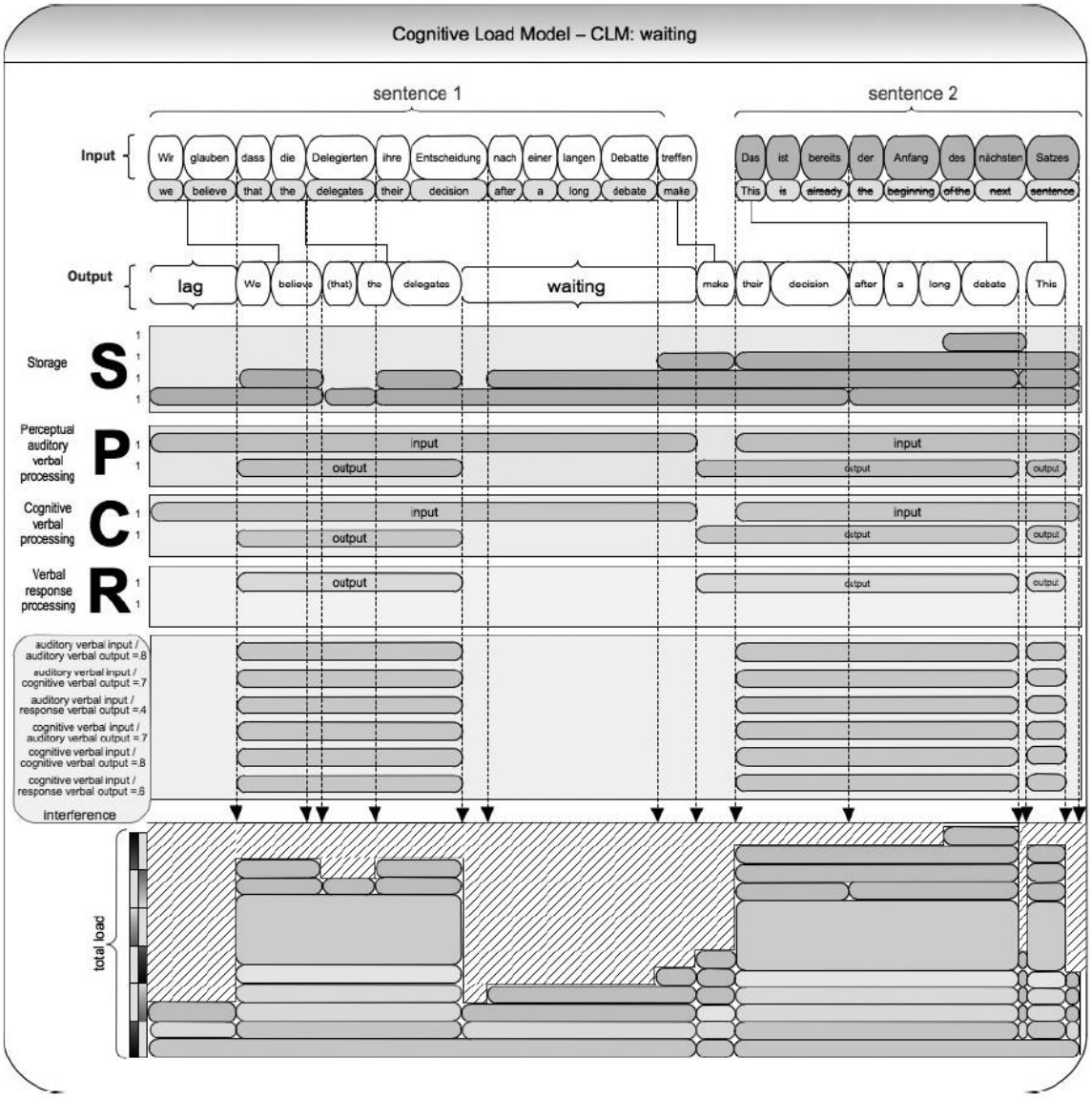



**Figure 2. 5** To be continued.

**B**

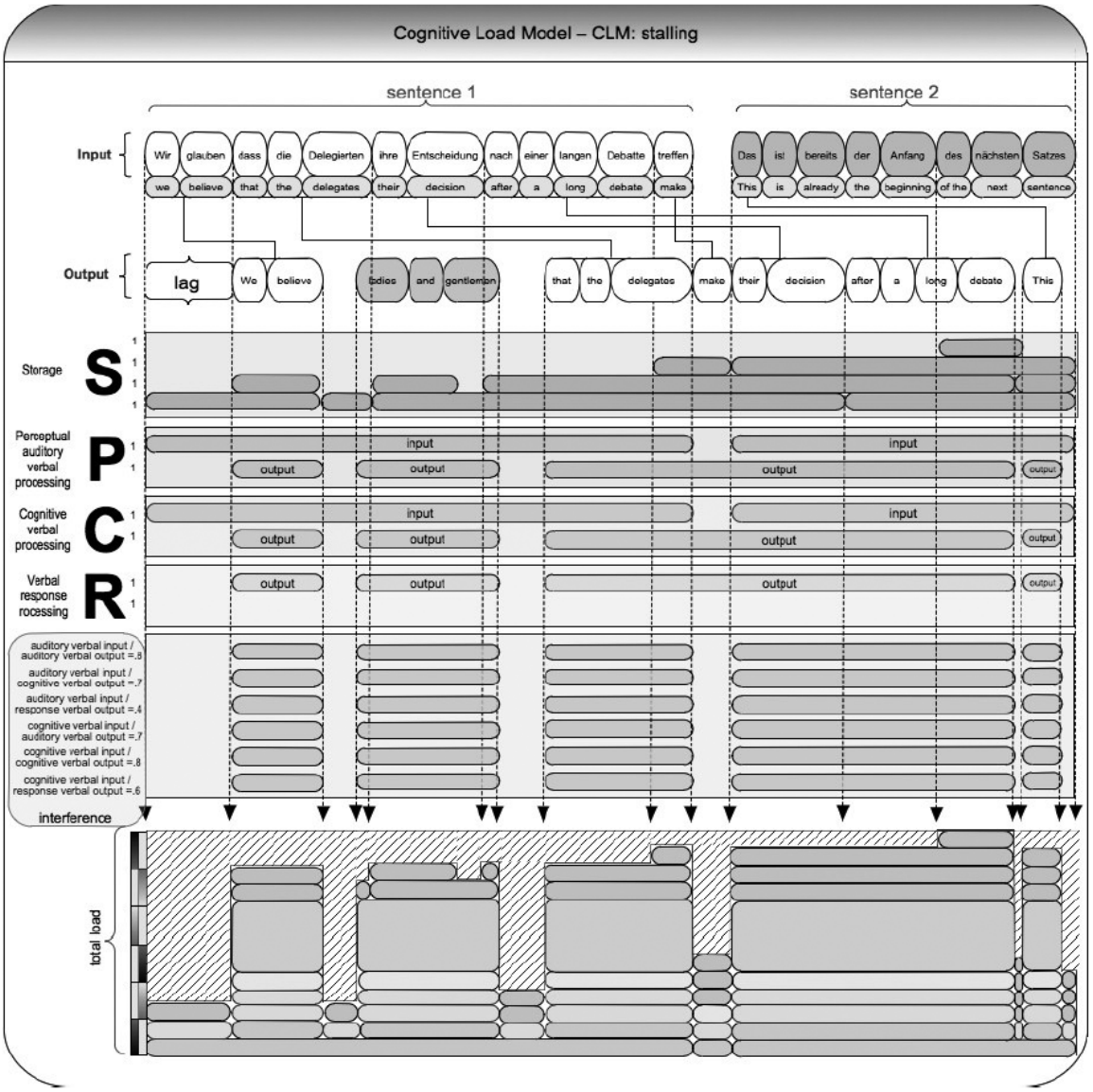



**Figure 2. 5** To be continued.

C

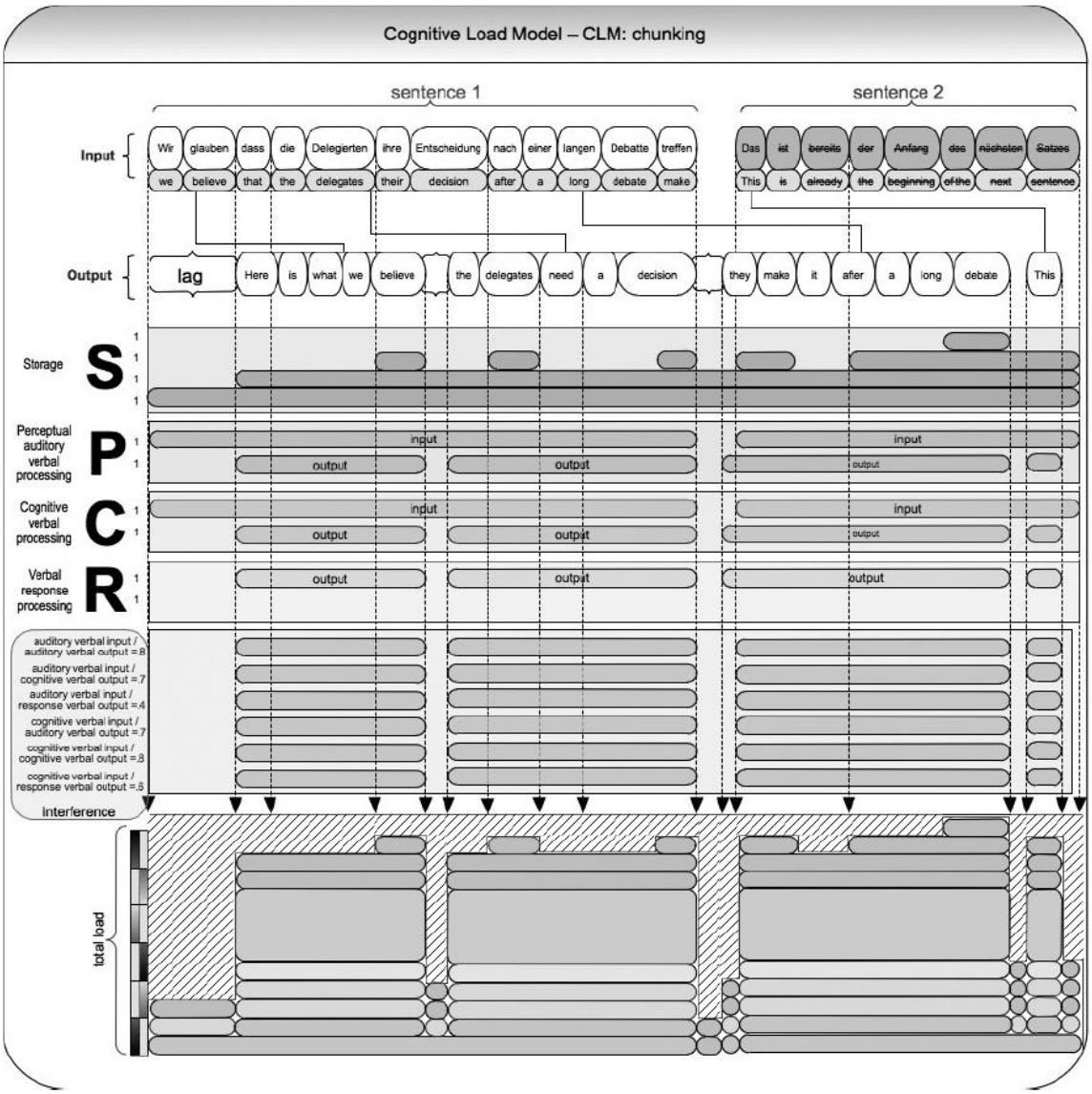



**Figure 2. 5** To be continued.

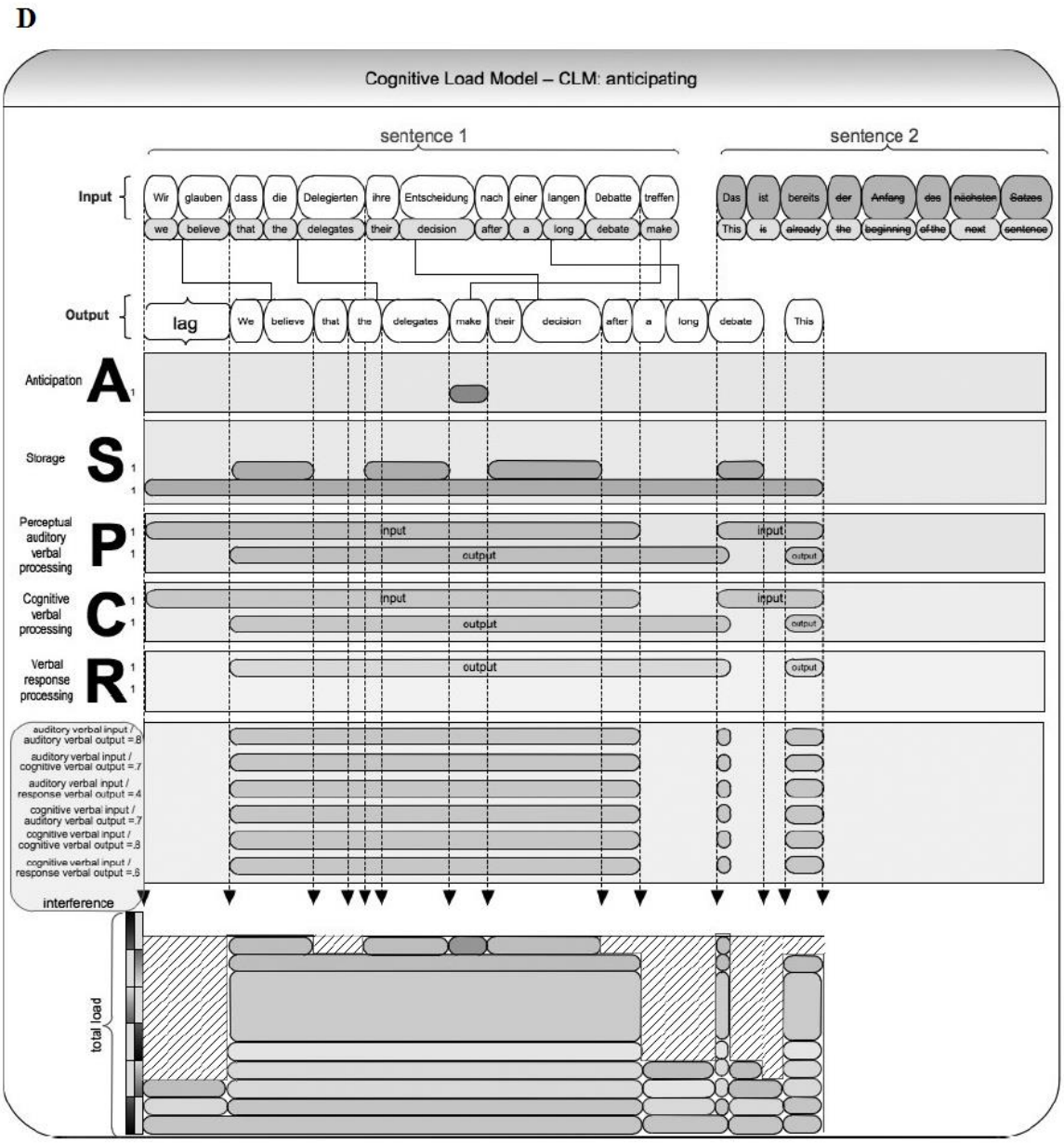



**Figure 2. 5** Cognitive Load Model for the *waiting* (graph A), *stalling* (graph B), *chunking* (graph C) and *anticipation* (graph D) strategy. Adopted from Seeber, K. G. (2011). Cognitive load in simultaneous interpreting: Existing theories—new models. *Interpreting*, *13*(2), 176-204.

The importance of anticipation for better SI performance is not only supported by Seeber (2011), but also by other researchers (Chernov, 2004; Gile, 2009; Van Besien, 1999). It is suggested that interpreters benefit from using their linguistic and extra-linguistic knowledge to predict the upcoming information in the SL, especially when the syntax of two languages (SL and TL) is asymmetrical (Seeber, 2001). Supporting evidence shows that interpreters are more accurate in anticipating verbs during SI (Seeber, 2001; Van

Besien, 1999), and that SI anticipation is not specifically confined to a certain language pair (Lederer, 1981, cited in Seeber, 2001).

The Cognitive Load model of Seeber (2011) is based on a fundamentally different point of view than the Effort model (Gile, 2009) in its assumption that instead of a single capacity resource pool, multiple resources are shared by all the concurrent tasks. If tasks demand the same processing resources, then more interference occurs than when they require resources from different pools. However, despite the differences of the two models, they can still be classified, in general, as capacity sharing models, since both of them share the assumption that concurrent tasks are performed by sharing resources that are limited in capacity. In the next section, existing psycholinguistic studies on SI are reviewed to provide a further understanding of the nature and components of processing that are involved in SI.

### 2.2.3 Psycholinguistic research on SI

Experiments have been conducted in various fields to explore the fundamental processing of SI. The measure of ear-voice span, the lag between the production of the SL information and the TL translation production, has been used to understand WM capacity and comprehension. The measurement of the ear-voice span is normally taken to be the duration of the lag or the number of words. Larger ear-voice spans have been shown to have a positive correlation with better SI performance, since the longer timeframe allows for sufficient information to be processed and avoid incorrect misunderstanding (Barik, 1975; but see Defrancq, 2015); however, a lag that is too long can exhaust limited WM capacity, which will result in greater text content omission and interpreting errors (Barik, 1975). It has been demonstrated that the ear-voice span is longer for paraphrasing (reformulating within the same language) than for interpreting (translating from one language into another), and shortest for shadowing (repeating the same sentence in the same language).

Additional evidence of the differences between the three tasks is provided by performance on recall (recalling the sentences from the message that needed to be interpreted, or paraphrased, or shadowed) and memory span (recalling a series of digits that are presented binaurally with the message) tests after interpreting, paraphrasing and shadowing. Gerver

(1974b) and Lambert (1988) found better recall after interpreting than shadowing. However, Darò and Fabbro (1994) present contradictory results, showing a larger digital span followed by shadowing than interpreting, while Christoffels and De Groot (2004) obtained similar recall results in the above three tasks under simultaneous conditions, and indicate that given the essential difference between these measures, it is inappropriate to train novice interpreters by practicing shadowing and paraphrasing to enhance SI performance (also see Moser-Mercer et al., 2000). Furthermore, Christoffels and De Groot (2004) also found relatively higher recall for all tasks under delayed conditions (which require participants to begin production after the SL sentence is finished) when compared to the simultaneous condition. Christoffels and her colleague (2004) account for the lower recall phenomenon in the simultaneous condition as a consequence of the interference of articulation with WM.

Articulatory suppression (AS) is the phenomenon that continuous irrelevant meaningless sound production during memory encoding interferes with retention in short-term memory (Baddeley, Lewis & Vallar, 1984). Since, as mentioned above, during SI, comprehension, memorising, and production coincide, it is important to investigate, as has been done in the past, whether AS will cause sustained interference for interpreters (Darò & Fabbro, 1994). The results are inconsistent on whether professional interpreters have advantages in preventing the interference of AS. Free recall tasks with AS, which requires participants to recall the list of words they have received when continuously producing as many sounds as possible, is usually adopted to explore AS by measuring the number of recalled words without order requirement. Numerous experiments have shown that compared to non-interpreters, and even high-span participants, professional interpreters are barely affected by AS (Injoque-Ricle et al., 2015; Köpke & Signorelli, 2012; Padilla et al., 2005) under both simple and complex conditions when required to produce a syllable (e.g. *ba*) or word (e.g. *one*) (Yudes, Macizo & Bajo, 2012). Explanations that have been proposed to account for the better performance of interpreters under AS indicate that either word knowledge plays an important role (Morales et al., 2015; Padilla et al., 2005), or that quick transformation from the loop to the buffer is the reason (Christoffels, 2006). However, results of no advantages for professional interpreters are also found in some experiments, which have shown that novice interpreters outperform professional interpreters (Liu et al., 2004; Köpke & Nespoulous, 2006). Since AS also involves WM, it is possible that WM differences are responsible for these mixed results.

According to Baddeley (2000), the definition of WM in psychology is that it is the limited capacity system where we can store information temporarily. WM has been widely agreed to play a critical role in SI (Christoffels, De Groot & Kroll, 2006; Gerver, 1975; Gile, 2009; Moser, 1978), even in bimodal (signed – spoken language) interpreting (Macnamara & Conway, 2014), and the working memory model that is favoured by most researchers on interpreting is the three-component model, which was first proposed by Baddeley and Hitch back in 1974 (for review, see Baddeley, 2000). However, the findings regarding whether professional interpreters outperform non-interpreters is a matter of some controversy. The majority of the findings have shown that professional interpreters have relatively larger memory capacity than non-interpreters by showing higher accuracy in the **dual N-Back task**, which asks participants to press the corresponding button when one of the concurrently presented stimuli (visual or auditory) is the same as the one N times ago (Morales et al., 2015); the **reading/listening span task**, which requires participants to recall all the last words of the presented sentences (Christoffels et al., 2003; Christoffels, De Groot & Kroll, 2006; Signorelli, Haarmann & Obler, 2011); the **non-word repetition task**, which asks participants to repeat after receiving the non-word over the headphone (Signorelli, Haarmann & Obler, 2011); and the (word or digital) **free recall task without AS** (Christoffels et al., 2003; Christoffels, De Groot & Kroll, 2006), even when participants' language proficiency has been controlled for (Christoffels, De Groot & Kroll, 2006). Christoffels and colleagues (2003) suggest that there is a correlation between WM and SI performance in non-trained bilinguals, indicating that participants with higher working-memory capacity can perform better SI. However, other experiments have failed to find advantages in professional interpreters (Chincotta & Underwood, 1998) and show that novice interpreting students outperform professionals in the listening span task (Köpke & Nespoulous, 2006). Signorelli and colleagues (2011) suggest that the inconsistent results might be due to the age impact as their data illustrated the advantages of younger interpreters, compared to professional interpreters and non-interpreters, in cued recall (the task that shows the first word of the six word list which has disappeared as a cued word, and requires participants to recall the other five words) and non-word repetition (the task that asks participants to repeat the non-word that was played over the headphone).

**Language production**, the process which is suggested to be allocated about 20% of the cognitive resources during interpreting while more than 80% of the resources are expended

on listening and comprehension (Gile, 2009), has also been investigated. Interpreters' production is an important aspect to explore since it is the information that audiences receive (that is, audiences who are not proficient in the language that the speaker is using). Disfluencies or pauses in the target output will influence the quality of the interpreting. While both professionals and students may have encountered the situation of hesitating or pausing during SI due to different reasons, such as being unable to produce the words or not being able to recall the previous message, professional interpreters have fewer and shorter un-natural pauses than interpreting students (Wang & Li, 2015). Such disfluencies often signal problems originating from language production (Bakti, 2009; Yudes et al., 2013) or WM. Other researchers have shown that when, as it happens occasionally, the speaker increases his speech rate, the interpreter usually adopts a strategy of omitting the redundant words and information (Chernov, 1979) by using shorter sentences with words that have fewer syllables (Sunnari, 1995, cited in Liu, 2008). The output of professional interpreters is also normally more logical and meaningful compared to that of novice interpreting students (Liu, 2008).

It is clear that professional interpreters provide higher quality production with fewer pauses than novice interpreters (Wang & Li, 2015). Furthermore, the output also shares, as suggested in the Effort model, the limited capacity, since language production in SI is not as easy as daily conversation and is definitely effortful (Gile, 2009). Thus, it is reasonable to assume that language production performance can, at least partially, interact with SI performance. Before reviewing the existing experimental results on language production, it is important to answer one question first, namely, how is language production processed? In other words, how is the SL turned into the TL?

Two interpreting strategies have been proposed and favoured: the **meaning-based strategy** and the **transcoding strategy**. The meaning-based strategy, as the name suggests produces chunks of information in the TL after receiving and understanding them (De Bot, 2000; Fabbro & Gran, 1994). That is, the language production in SI is similar to daily spontaneous speech production (see Fig. 2.6). In contrast, the transcoding strategy suggests that there are some links between the two languages and the interpretation can be conducted even without comprehending the SL message first (see Fig. 2.6; Paradis, 1994; also see Section 2.2.2.1). That is, the language production in SI is not the same as normal language production because the transcoding does not take place at conceptual (meaning)

level. Paradis (1994) suggests that the former strategy is adopted by interpreting students while the latter one is favoured by professional ones, while others suggest that professional interpreters might adopt the meaning-based strategy more often while using the transcoding strategy to handle a message that is difficult to understand (Darò, 1994). The meaning-based strategy is favoured by some researchers with the supporting evidence that professional interpreters outperform novice interpreters in detecting semantic errors but not syntactic errors (Fabbro, Gran & Gran 1991; also see Yudes et al., 2013) and recall less sentence form (Isham, 1994), suggesting that interpreters are more sensitive to the meaning of the message instead of its syntax as a consequence of the interpreting strategy they adopted.

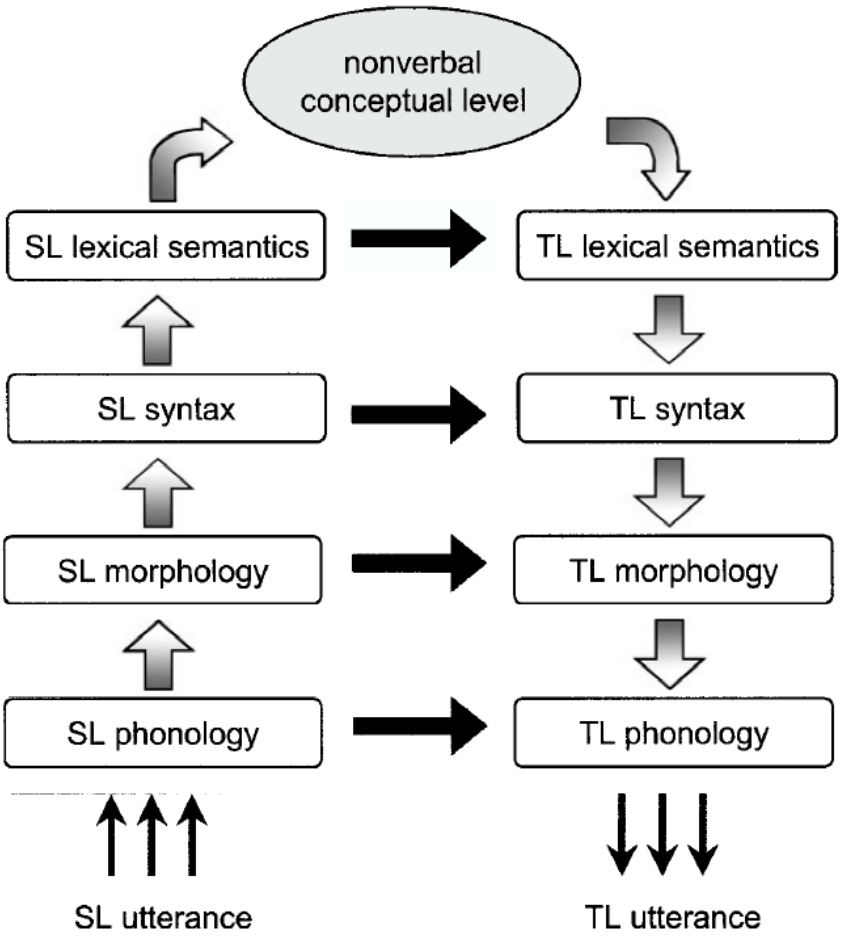



**Figure 2. 6** The model of simultaneous interpreting strategies based on Paradis (1994). The big grey arrows represent the meaning-based strategy while the narrow black arrows indicate the transcoding strategy. From Kroll, J. F., & De Groot, A. M. (Eds.). (2005). *Handbook of Bilingualism: Psycholinguistic Approaches*. Oxford: Oxford University Press. p. 460.

Although no definite results have been provided to support either of these two interpreting strategies, it is clear that the meaning-based strategy seems more difficult to rule out. Thus, it is worth exploring whether interpreters are different from non-interpreters in word production, which is suggested to be the basic unit of sentence production.

Surprisingly, relatively few studies have explored the difference between professional interpreters and bilinguals in lexical (word) retrieval (Christoffels et al., 2006; Lijewska & Chmiel, 2014). Christoffels and colleagues' (2006) results showed no difference between professional interpreters and highly proficient bilinguals in language production, when using a picture naming task and word translation task in both languages. Similar results were found by Lijewska and Chmiel (2014), indicating no difference in lexical access between trilingual interpreting trainees and non-trained trilinguals in translation tasks which required participants to translate from their third language (non-dominant language) to the first and second languages (dominant languages). Ibáñez and colleagues (2010) have compared the performance of professional translators and carefully matched bilinguals in a reading sentence task, but they found mixed results in sentence reading latencies. Apart from these few inconsistent experiments, there is one interesting study which explored the relationship between SI performance and picture naming and word translation latencies in untrained bilinguals (Christoffels et al., 2003), showing that better performance is correlated with shorter response latencies.

One of the few interpreting models that provides details on language production is De Bot's (1992, 2000) model which was inherited from the word production model of Levelt and colleagues (1999), suggesting that interpreters adopt the meaning-based strategy and they share the same output processes as bilinguals do during SI. Furthermore, a larger number of research results have shown that bilinguals, even proficient ones, take a longer time to produce a word than monolinguals even in the dominant language (Ivanova & Costa, 2008). The question that arises is whether this is also the case for professional interpreters. The following sections discuss the basic concepts of language production of monolinguals and bilinguals before further exploring whether professional interpreters are different from non-interpreters, both bilinguals and monolinguals, in language production.

## 2.3 Lexical access of bilinguals and monolinguals

What happens when a person produces words and sentences? Is there any empirical way to explore how language production works, even if it is a mental process? These questions are fundamentally important since language production is central to achieving the goal of communication, and has been explored in psycholinguistics, neurolinguistics and cognitive neuropsychology. In order to understand how language production works, several methods have been used, including the performance of speech errors, and the reaction time in naming and lexical decision tasks, for both brain-damaged subjects (Coughlan & Warrington, 1981; Dell et al., 1997; Foygel & Dell, 2000; Hillis, Rapp, Romani & Caramazza, 1990; Le Dorze & Nespoulous, 1989) and normal subjects (Caramazza, Costa, Miozzo & Bi, 2001; Cutting & Ferreira, 1999; Jescheniak & Levelt, 1994; Shatzman & Schiller, 2004; Starreveld & La Heij, 1995, 1996).

Questions about language production become even more complex when bi-/multilingualism is taken into consideration. The fact that monolinguals can speak only one language, while bilinguals can speak two or more different languages, raises the question of whether there is any difference between the language production processes of monolinguals and bilinguals. More complex questions also arise. Is there any difference between unbalanced bilinguals, who are not as proficient in their second language, and balanced bilinguals (who are equally proficient in their two languages), such as professional simultaneous interpreters during language production?

To further explore this question, this section focuses on the existing literature on monolingual and bilingual language production, specifically, on word production. Section 2.3.1 first summarises the general structure and principles that most language production models have adopted, and outlines each stage involved during word production. Section 2.3.2 illustrates the main methods that have been widely used to explore speech production, and the main findings of such research, particularly to provide support for the notion that there are multiple distinct stages in language production. Section 2.3.3 introduces two influential theories of speech production which have been used to account for these findings. Following this, Section 2.3.4 reviews experiments exploring bilinguals' lexical access, after which Section 2.3.5 briefly outlines the most influential theories of

bilinguals' lexical access models. Section 2.3.6 concludes with a comparison of the difference between monolinguals' and bilinguals' word production.

### 2.3.1 Basic principles of word production

Generally speaking, all theories of word production that offer an account of the selection and retrieval of single isolated words share the opinion that multiple sequential stages are involved during word production, and the representations with the highest activation level in each stage will be selected. Theories of word production differ from one another in a variety of ways, as will become evident in the subsequent discussions. A general word production model which is compatible with most theories is set out in Ferreira and Pashler (2002), and illustrated in Figure 2.7.

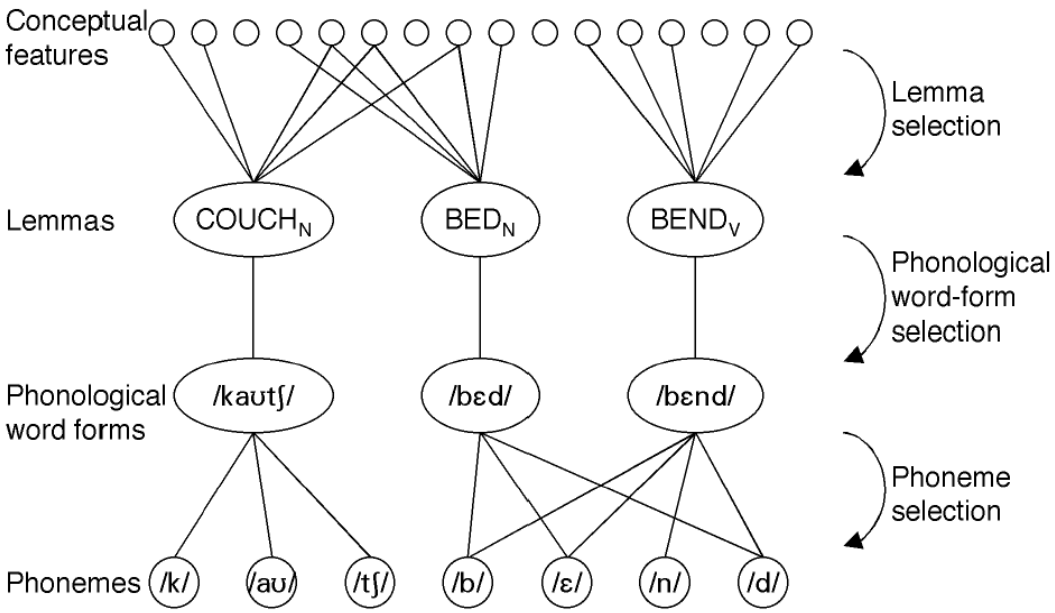



**Figure 2. 7** A model of part of the word production lexicon. Information flows from top to bottom. Adopted from Ferreira, V. S., & Pashler, H. (2002). Central bottleneck influences on the processing stages of word production. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 28(6), 1187-1199.

The starting point of word production is the **conceptual (or semantic) stage**, which involves the activation of the information or meaning that needs to be expressed. Regardless of whether the view of conceptual/semantic representation is non-decompositional (Garrett, 1982; Levelt et al., 1999; Starreveld & La Heij, 1996), or decompositional (Caramazza, 1997; Dell, 1986; Katz & Fodor, 1963; Osgood, 1963), there

is agreement that a "group" of conceptually related words will be activated. Non-decompositional views of conceptual/semantic representation hold that the semantic concept for an intended word is a single indivisible unit, which cannot be separated into multiple features. For example, the concept of the word *owl* is OWL, but not BIRD or CAN FLY or NOCTURNAL, etc. In contrast, decompositional or componential views of conceptual/semantic representation suggest that multiple semantic features (e.g. BIRD, CAN FLY, NOCTURNAL, etc.) constitute the concept for an intended word (*owl*).

Subsequent to the conceptual/semantic stage, activation then spreads to the **lemma stage**. The term "lemma" (also called a "lexical representation") was first used in this context by Kempen and Huijbers (1983), who defined it as a lexically specific representation; in other words, a lemma represents a word. Selection of the appropriate specified word is required, since unintended semantically related lemmas will also be activated. The retrieval of a lemma is influenced by a few factors which are accepted by nearly all models of word production:

- **The number of lemmas that can express the same concept**. The smaller the number is, the easier the retrieval will be (Lachman, 1973).
- **The constraint of sentence context**. The higher the predictability of the word is in the context of the sentence, the faster selection of the lemma will be (Griffin & Bock, 1998).
- **Whether nouns are animate or not.** Evidence shows that animate nouns (referring to a class of creatures that have life, like persons and animals, e.g. *dog*) are easier to retrieve than inanimate nouns (referring to a category of things and concepts, e.g. *desk*) (Rohrman, 1970).
- **Whether the lemma is concrete or not.** Whether the word arouses a concrete image is also correlated with lemma retrieval, suggesting that concrete words (e.g. *dog*) are easier to recall than abstract words (e.g. *spirit*) (Paivio, 1971; Van Hell & De Groot, 1998b).

The next stage in the multistage process is the **lexeme stage,** where the activation spreads to the (phonological) word-form or phonological representation, which includes the whole sound segment of the selected lemmas, the order of those segments, and the word's morphological properties. The final stage which receives activation is the **phoneme**

**(segment retrieval) stage**. Phonemes are the individual speech sounds that make up the word; for example /p/, /l/, /ɑ:/, /n/, /t/ are the phonemes that make up the word *plant*. Each individual sound composing the to-be-produced word is retrieved and organised into syllables during this stage. In sum, after determining the message, **lemma selection** is the first selection step during word production, followed by **phonological word-form selection**. The final selection stage is **phoneme selection,** before the speech motor stage commences.

These basic principles form the foundation for models of spoken language production at lexical level. These models, and particularly the similarities and differences between them, are discussed in more detail in Section 2.3.3. Before turning the attention to these models, however, the next section first considers some of the methods that have been used to investigate language production, and the main empirical findings from studies using these methods.

### 2.3.2 Investigating monolingual speech production: Main methods and findings

From the moment a person intends to produce a word until he/she is at the point of articulating the sound of the word, all the processes of retrieving and selecting the target word, as outlined in Section 2.3.1, are mental activities. The question is how word-production processes can be explored, when these are mental processes. In this section, several methods that are commonly adopted as measurements to explore language production are introduced, including speech errors, the tip-of-the-tongue phenomenon, and the picture-word interference paradigm.

#### 2.3.2.1 Speech errors

Although the speech that people produce is usually accurate and understandable, it commonly happens that people produce the wrong word (e.g. *spoon* instead of *fork*) or sound (e.g. *hat* instead of *cat*) during speech production, whether spontaneous or pre-prepared. These kinds of mistakes are called speech errors.

The analysis of speech errors is one of the most reliable ways to gain insight into language production, and has been used extensively by many researchers (Boomer & Laver, 1968; Butterworth, 1981; Dell, 1990; Dell & Reich, 1981; Fay & Cutler, 1977; Fromkin, 1971, 1973; Garrett, 1975, 1976, 1988; Stemberger, 1985; Shattuck-Hufnagel & Klatt, 1979) to try to reveal the mystery of how the cognitive system works during language production. Speech errors are either collected by monitoring people in natural communication contexts (Berg, 1992), or are elicited in experimental design tasks, which require subjects to complete sentences (Schriefers & Jescheniak, 1999).

There are several types of naturally occurring speech errors that arise during normal spontaneous speech. The most commonly occurring errors include semantic substitution errors, morpheme substitution errors, and sound substitution errors. **Semantic substitution errors** occur when a speaker mis-produces a semantically related word instead of the word intended to be produced, such as producing the word *flower* instead of the word *plant*. This suggests that groups of semantically related words (words that share at least some part of the intended meaning), are activated simultaneously. Speakers make **morpheme substitution errors** during language production by replacing a morpheme (the smallest meaningful unit, such as *key* and *board* in *keyboard*) in a word. For example, if a speaker says *useLESS* instead of *useful* (see Griffin & Ferreira, 2006), a morpheme substitution error has occurred. The last common type of error is **sound substitution errors.** The substitution of an individual sound or phoneme, like the /l/ in LAUGH, by another segment, such as /h/, represents this type of error. Thus, the word LAUGH /la:f/ can be mis-produced as the similar-sounding word HALF /ha:f/ even though the two words are not semantically related.

These different types of speech errors provide strong evidence that multiple stages are involved during language production. Most importantly, speech-error data show that there are separate stages involved: the stage that involves the meaning of the word (semantic stage), and the stage that involves the sounds and the combination of sounds of the word (phonological stage). These stages must be separate since sound substitution errors are normally made by mispronouncing a word with similar sounds but unrelated meanings (e.g. mis-producing *cat* as *hat*), while semantic substitution errors are made by unintentionally producing a semantically related word within the same conceptual group (e.g. mis-producing *cat* as *dog*).

If it is assumed that there are multiple stages during language production, a subsequent question would be what the sequence of these stages is. The so-called tip-of-the-tongue (TOT) phenomenon offers a way to explore this question.

#### 2.3.2.2 Tip-of-the-tongue (TOT) phenomenon

The TOT-phenomenon occurs when a person momentarily finds the pronunciation of a word inaccessible, even though they are certain its meaning is stored in the memory (Brown, 1991). This frustrating experience has been explored by many researchers in different languages (Biederman et al., 2008; Gollan & Silverberg, 2001; Kikyo et al., 2001; Vigliocco et al., 1997) across the lifespan (Wellman, 1977; Burke et al., 1991). This research converges on the finding that TOT is a universal experience which is influenced by age and the frequency of the target word (Brown, 1991).

The explanation that most theories of language production offer to account for the TOT-phenomenon is that it is the consequence of a failure to retrieve the phonological representation of a word, but successfully accessing its lexical representation. Since the speaker knows the meaning of the word but is temporarily unable to pronounce it, the lexical information must have been retrieved, suggesting that this stage must occur prior to retrieving the pronunciation (or phonological representation of a word). Based on these findings, most models of language production (discussed in more detail in Section 2.3.3) agree that the multiple stages involved in lexical access are sequential, and the semantic stage precedes the phonological stage.

Furthermore, evidence from the investigation of normal subjects and aphasic patients' performance indicates that although subjects are in a TOT-state, they can still retrieve the initial phoneme or letter and grammatical gender of the word (Badecker, Miozzo & Zanuttini, 1995; Caramazza & Miozzo, 1997; Henaff Gonon et al., 1989; Goodglass et al., 1976; Kay & Ellis, 1987; Miozzo & Caramazza, 1997; Vigliocco, Antonini & Garrett, 1997). Based on this, it has been proposed that access to syntactic features also precedes the phonological stage.

While contributing to our understanding of language production, data from speech errors and the TOT-phenomenon cannot, however, provide a measure to explore the time latency when producing a word, or provide any measures to investigate the time course of each stage involved during language production. To collect this kind of information, particularly on the time course of word production, the picture-word interference paradigm has most commonly been used.

#### 2.3.2.3 The picture-word interference paradigm

The picture-word interference paradigm is one of the most widely used methods to explore language production in psycholinguistics (Costa, Mahon, Savova & Caramazza, 2003; La Heij, Mak, Sander & Willeboordse, 1998; Schiller & Caramazza, 2002; Schriefers, 1993; Schriefers, Jescheniak & Hantsch, 2005; Spalek & Schriefers, 2005; Jescheniak & Levelt, 1994; Kempen & Huijbers, 1983; Levelt et al., 1991; Schriefers, Meyer, & Levelt, 1990). In this paradigm, subjects are usually presented with a picture with a distractor word imposed on it (visually), or played simultaneously (auditorily), and asked to name the picture as quickly as possible while ignoring the word. The type of distractor word and the SOA between the picture and distractor word are often manipulated to achieve the goal of exploring the processes of word production.

A large number of experiments in this paradigm have shown that when semantically related distractors (e.g. the word *spoon* as a semantically related distractor of the picture FORK) are presented simultaneously with or precede the presentation of the picture by up to 400 ms, the response latency to the picture naming is prolonged, compared to an unrelated control word. This effect is called semantic interference (Glaser & Düngelhoff, 1984; Lupker, 1979; Roelofs, 1992; Starreveld & La Heij, 1995, 1996).[2] When the semantically related distractor word is presented after the picture appears, the response latency is (mostly) either barely affected or speeded up, relative to the control condition. One exception is evident in Janssenetal et al. (2008), who obtained a semantic interference effect even when the distractor is presented 1000 ms after the target picture. Based on this,

[2] The effect of semantically related distractors in a different language to the picture-naming response will be discussed in Section 2.3.4.2, dealing with the speech production of bilinguals.

the authors argue that semantic interference occurs at the post-lexical stage but not at the lexical stage. However, a study by Mädebach et al. (2011) fails to replicate these findings.

One explanation offered to account for semantic interference is that a semantically related distractor receives activation from two sources, namely the distractor itself and the intended word, since they are semantically related; whereas an unrelated distractor only receives activation from itself. Therefore, the activation of semantically related distractors is stronger than is the case for unrelated words, and this leads the semantically related distractors to be more competitive to the intended word, compared to unrelated distractors. This account is referred to as the **lexical selection by competition hypothesis** (La Heij, 1988; Bloem & La Heij, 2003; Levelt et al., 1999), and is based on the assumption that the more activation a distractor word receives, the more competitive the lexical selection will be.

However, an alternative explanation for the semantic interference effect is the **response exclusion hypothesis** (Finkbeiner & Caramazza, 2006; Finkbeiner, Gollan & Caramazza, 2006; Mahon et al., 2007). The key idea of this hypothesis is that the semantic interference effect, which arises at post-lexical stage, is correlated with the time that the decision mechanism takes to exclude the distractor words based on the response-relevant criteria. The response-relevant criteria involve the degree to which the distractor word meets the general semantic constraint demanded by the target word. The semantically related distractor satisfies more of the intended word's response-relevant criteria compared to the unrelated distractor, and consequently, the latency to exclude the semantically related distractor is longer than for the unrelated distractor.

Interestingly, if the distractor is a word which is related to the meaning of the target picture, but is not conceptually related (e.g., the word *bone* is associated with *dog*, but it is not conceptually related to *dog* like the word *cat*), then the response latency to the target picture will not be prolonged but facilitated, compared to the unrelated control distractor (Mahon et al., 2007). Similarly, if the distractor word belongs to part of the target picture (e.g. the word *ink* refers to part of the *pen*) (Costa, Alario & Caramazza, 2005), a semantic facilitation effect can also be obtained. Furthermore, when the distractor is a within-category semantically related word, the response latency is faster when the distractor is semantically closer to the intended word (e.g. *zebra* to the picture HORSE), compared to

when the distractor is semantically far from the target word (e.g. *whale* to the picture HORSE) (Mahon et al., 2007). These semantic facilitation results are consistent with the prediction of the response exclusion hypothesis, arguing against the selection by competition hypothesis, because an interference effect should be expected instead of a facilitation effect based on the selection by competition hypothesis.

In contrast, if a phonologically related distractor word (e.g. the word *doll* as the distractor of the picture DOG) precedes the picture that needs to be named, results are inconsistent. Some experiments have shown that the effects of phonological distractors are absent at negative SOA (when the phonologically related distractor word is presented earlier than the picture) (Jescheniak & Schriefers, 1998; Schriefers, Meyer & Levelt, 1990). Other studies, however, have reported that response latency is facilitated by phonological effects at early SOA (Damian & Martin, 1999; Meyer & Schriefers, 1991; Starreveld, 2000).

In sum, the picture-word interference paradigm has shown that compared to unrelated control words, semantically related distractor words compete for selection with the target word and consequently prolong the response latency. Phonologically related distractor words, in contrast, facilitate the retrieval of the target word's phoneme, and therefore shorten the latency to produce the intended word. Moreover, the results of picture-word interference tasks have also illustrated the different time course of semantic and phonological effects when manipulating SOA, demonstrating that the semantic stage occurs prior to the phonological stage since a semantic effect occurs when the distractor word is presented prior to or simultaneous with the picture, and a phonological effect occurs when the distractor word is presented after the picture.

The above methodologies have provided many of the core insights about lexical access. Based on empirical evidence from many studies using the above methodologies, there is a general consensus in theories of word production that a lexical/lemma selection stage and a phonological/lexeme selection stage can be distinguished (Bock, 1982; Bock & Levelt, 1994; Burke, MacKay, Worthley & Wade, 1991; Butterworth, 1989; Dell, 1986; Dell & O'Seaghdha, 1992; Dell, Schwartz, Martin, Saffran, & Gagnon, 1997; Fay & Cutler, 1977; Fromkin, 1971; Garrett, 1975, 1980; Harley, 1984; Kempen & Huijbers, 1983; Levelt, 1993; Roelofs, 1992; Stemberger, 1985). However, this is where the broad agreement among theories of word production ends, and current theories offering an account of

speech production differ in significant ways. The following section outlines these differences, and summarises two of the most influential current theories of monolingual language production.

### 2.3.3 Theories of word production

Language production research has a long history, and the systematic study of word production based on spontaneous speech error corpora can be traced back to the 1960s (e.g. Boomer & Laver, 1968; for review see Levelt, 1999). The first computational word production model based on speech error data was proposed by Dell (1986). Since then, several models that account for word production have been proposed (Levelt et al., 1999; Roelofs, 1997). All the models agree that multiple stages are involved in word production, and that activation spreads from the conceptual stage to the phonological stage. However, models of word production differ from each other in various details. The contention is essentially focused on two issues: the way that activation flows, and the number of stages involved in lexical access.

In this section, the two most influential models accounting for lexical access are introduced: WEAVER++ (Levelt et al., 1999) and the independent network model (Caramazza, 1997; Caramazza & Miozzo, 1997, 1998). Both of these models offer satisfactory accounts of a large number of existing research results. However, the two models hold different views regarding the above-mentioned points of controversy: language production interactivity, and number of stages. Levelt and colleagues (1999) favour the assumption that four stages are involved in word production before the commencement of articulation: the conceptual stage, lexical stage, phonological word-form stage and phoneme stage (also see Cutting & Ferreira, 1999; Roelofs, 1992). However, Caramazza (1997) argues that there is no distinction between the lexical and phonological word-form stages, suggesting that only three stages are involved in word production: the conceptual stage, the lexeme stage (the stage where both lexical and phonological word-form nodes exist), and the phoneme stage.

The WEAVER++ model of Levelt et al. (1999) favours the assumption that activation flows in a **discrete** way, suggesting that the word production process is strictly successive, and activation cannot pass down to the subsequent stage (e.g. the phonological stage) until

the completion of the previous stage (e.g. the lexical stage) (also see Kempen & Huijbers, 1983; Levelt et al., 1991; Schriefers et al., 1990). In this case, the activity in the subsequent stage cannot influence the previous one, and the later stage can only be influenced by the target lemma in the earlier stage. In contrast, the independent network model favours the assumption that the word (production) process is characterised by **cascaded** activation. In other words, activation flows freely through all stages and later stage(s) can be activated before the completion of the previous one (also see Dell, 1985; Dell & O'Seaghdha, 1991). This means that the later stage representations can also be activated even though they correspond to the unintended word. Therefore, the later stage can be influenced by the unintended nodes in the earlier stage.

#### 2.3.3.1 WEAVER++ (Levelt, Roelofs & Meyer, 1999)

One of the crucial traits of the model of language production proposed by Levelt et al. (1999) (and which sets it apart from the independent network model (Caramazza, 1997; Caramazza & Miozzo, 1997, 1998) discussed in Section 2.3.3.2), is that it favours the assumptions of non-decompositionality (see Section 2.3.1), a discrete activation spreading process (also see Roelofs, 1992), and feedforward activation only. The assumption of discrete activation is supported by substantial evidence showing that the semantic interference effect is absent at positive SOA (when the semantically related distractor appears later than the picture) in the picture-word interference task, suggesting that the processing at lexical stage is completed when the distractor is presented (see Section 2.3.2). Thus, no semantic interference should be obtained in this condition (Levelt et al, 1991; Schriefers et al., 1990).

This model distinguishes three strata: the conceptual, lemma and form strata. Nodes in the conceptual stratum, lemma stratum, and form stratum represent lexical concepts, lemmas with their syntactic properties, and word-form and phoneme segments, respectively (see Fig. 2.8). Speech production starts with conceptual preparation, the stage that precedes the lexical/semantic stage. Levelt et al. (1999) make it clear that conceptual preparation is necessary "leading up to the activation of a lexical concept" (p. 3). After activating a lexical concept representation, activation then spreads to other concept nodes via IS-TO links and activate semantically related lexical concept representations (see Fig. 2.8). The notion of the IS-TO link was first proposed by Roelofs (1992, 1993), and can be explained

by the connection *as OWL (X, Y) is to BIRD (X, Y)*. A group of words are activated as a consequence of these conceptual connections, and therefore, the selection of the specific word is needed.

Lexical selection, or lemma selection, involves retrieving a single word, which can express the given lexical concept, from the mental lexicon. Lemma selection is correlated with the amount of activation, such that the lemma with the highest degree of activation will be selected. As mentioned above, one of the basic assumptions of this model is that the lexical concept is non-decompositional. Therefore, the word OWL will be retrieved since it receives the full activation from concept node OWL and becomes the most highly activated representation, while the lemma BIRD only receives a proportion of the activation. Moreover, the selection of syntactically driven function words also occurs during this stage. Function words have little or no lexical meaning (e.g. *that* in the sentence "the word *that* is presented on the screen…." is a function word that has no lexical meaning, thus, it is syntactically driven but not lexically driven), and play a role in completing the correct grammatical structure of the sentence.

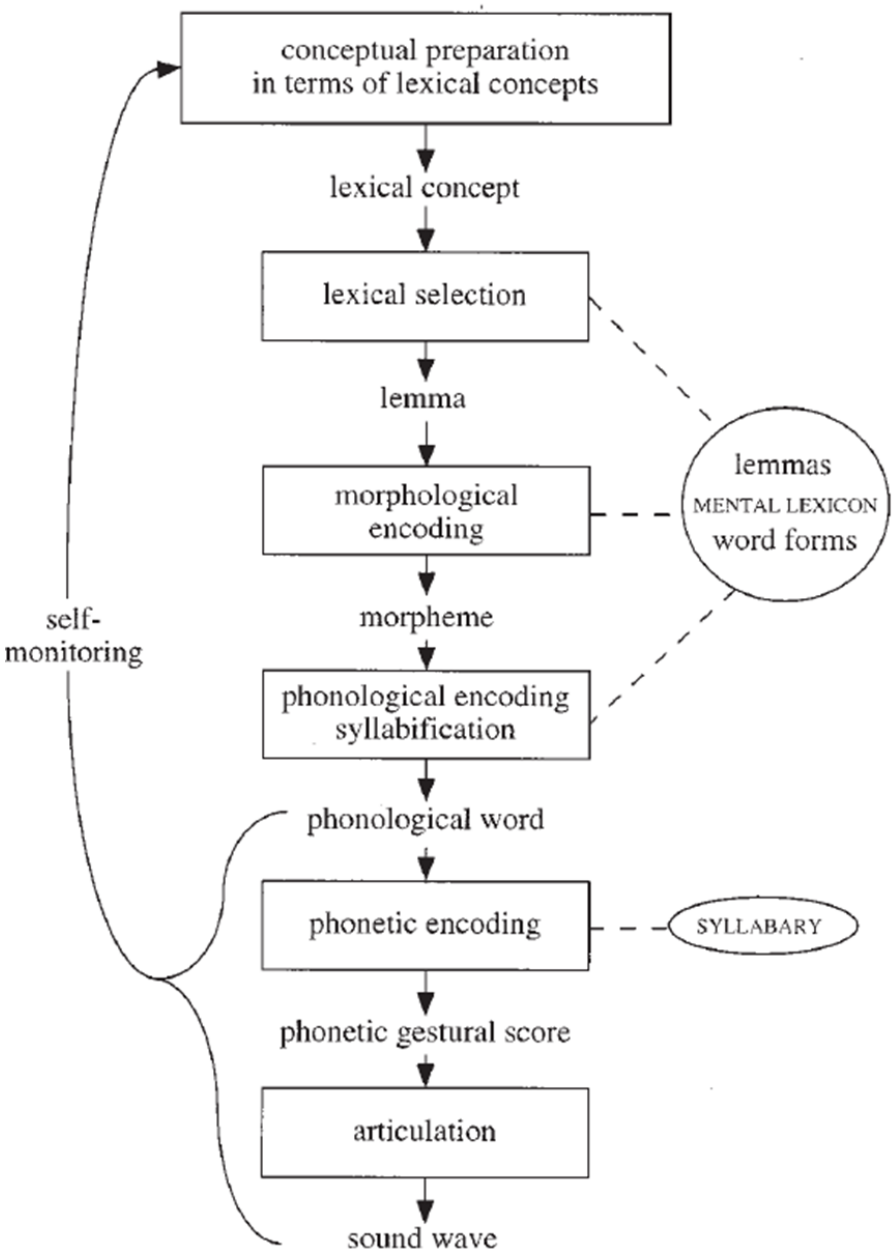



**Figure 2. 8** Outline of the Levelt et al. (1999) theory of language production. The procedure starts with conceptual preparation, before continuing to lexical selection, morphological and phonological encoding, and phonetic encoding, with the articulation stage as the endpoint. Self-monitoring of the output is also included in this theory. Adopted from Levelt, W. J., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences*, *22*(01), 1-38.

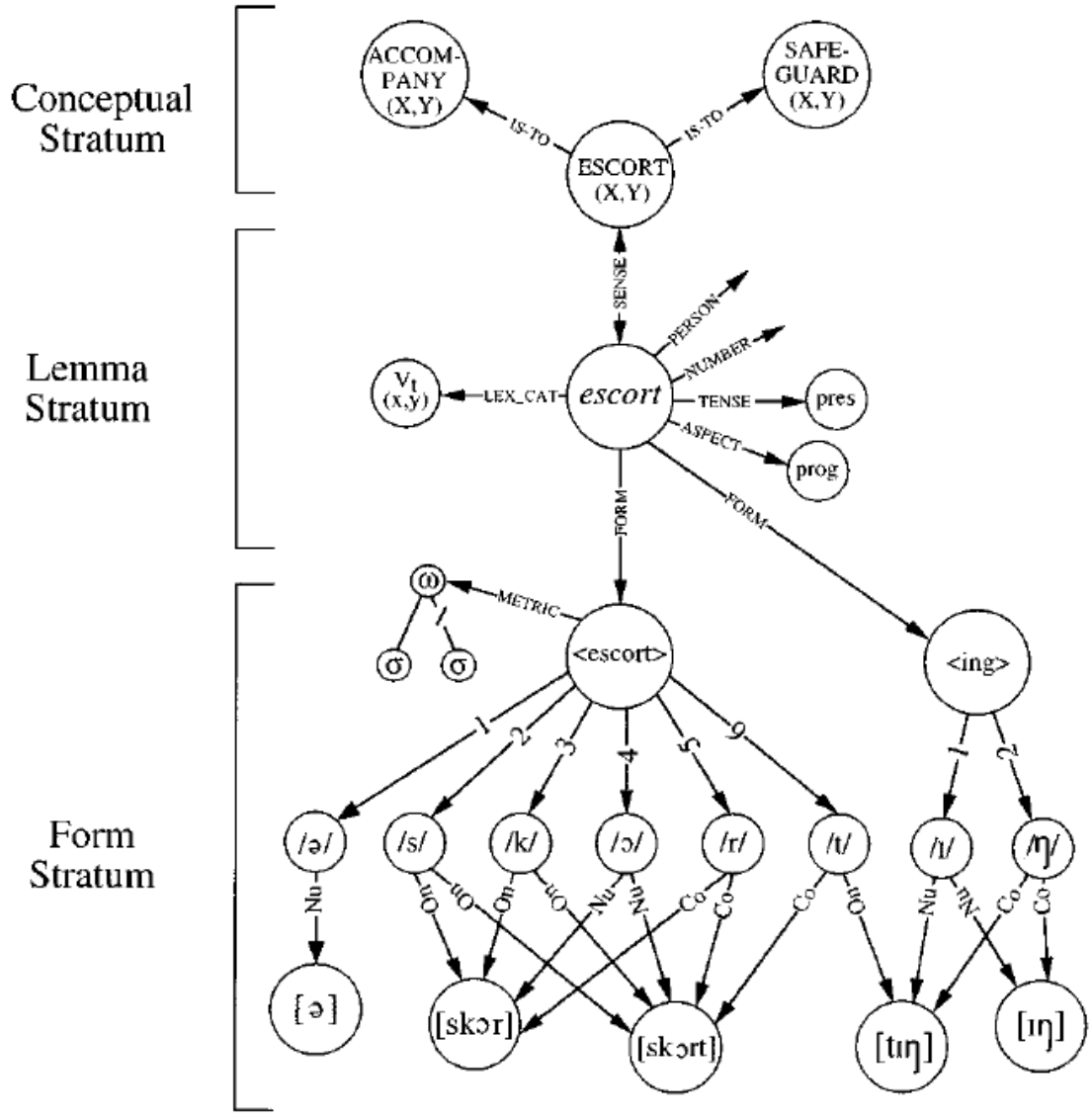



**Figure 2. 9** The structure of the lexical network during language production. Nodes in conceptual stratum, lemma stratum, and form stratum represent lexical concepts, lemmas with their syntactic properties, and word-form and phoneme segments. Adopted from Levelt, W. J., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences, 22*(01), 1-38.

Once lexical selection is complete, the activation passes down to the intended lemma. In this respect, it is important to mention that syntactic features are part of the properties of the lemma. Levelt (1989) proposed this assumption, based on evidence from brain-damaged patients which shows that subjects know the grammatical gender of a word they cannot access to produce (Badecker et al., 1995).

The next step is retrieving the correct phonological shape of the selected lemmas, which can be termed the (phonological) word-form. In contrast with the independent network model (Caramazza, 1997) (discussed in more detail in Section 2.3.3.2), Jescheniak and Levelt (1994) propose that the phonological word-form stage is distinct from the lemma stage. The most compelling evidence to support this assumption comes from the existence of homophones. Homophones share the same pronunciation (sometimes even the same orthography), but they do not share the same meaning. For example, the word *bat* is a phonologically and orthographically identical homonym, which can refer to either a kind of animal or a piece of equipment used in sports like baseball or cricket to strike a ball. Moreover, homonyms can also have different syntactic features; for example, one could be a noun, and the other a verb or adjective. Therefore, the level that distinguishes homonym pairs must be the semantic level (since they have different meanings) but not the level of phonological word-forms (since they have identical pronunciation).

Further support for the existence of separate lemma and phonological word-form stages comes from the word frequency effect. The frequency of a word is correlated with the response latency to name this word (Oldfield & Wingfield, 1965), and the more frequently a word is encountered in daily life, the more quickly the word can be accessed. The word frequency effect exists also for the word-translation task (De Groot, 1992). The word frequency effect is indicated as located at the phonological word-form stage and mainly affects the retrieval of word-form (Dell, 1990; Jescheniak & Levelt, 1994). This assumption can also be supported by evidence from studies investigating the homophone frequency effect. The high-frequency word *week* is the homophone of the low-frequency word *weak*, and the non-homophone control word (e.g. *moon*) matches the sum frequency of the words *week* and *weak* (Jescheniak & Levelt, 1994; Jescheniak, Meyer & Levelt, 2003; but see Caramazza, Costa, Miozzo & Bi, 2001). Previous evidence has shown that the response latency of a low-frequency word is longer than the latency of a high-frequency word. Therefore, it is reasonable to infer that the low-frequency word *weak* "benefits" from its homophone, the high-frequency word *week*, at phonological word-form selection stage (see Fig. 2.10), which is distinct from the lemma stage since both *week* and *weak* share the same word-form /wi:k/ but have different meanings.

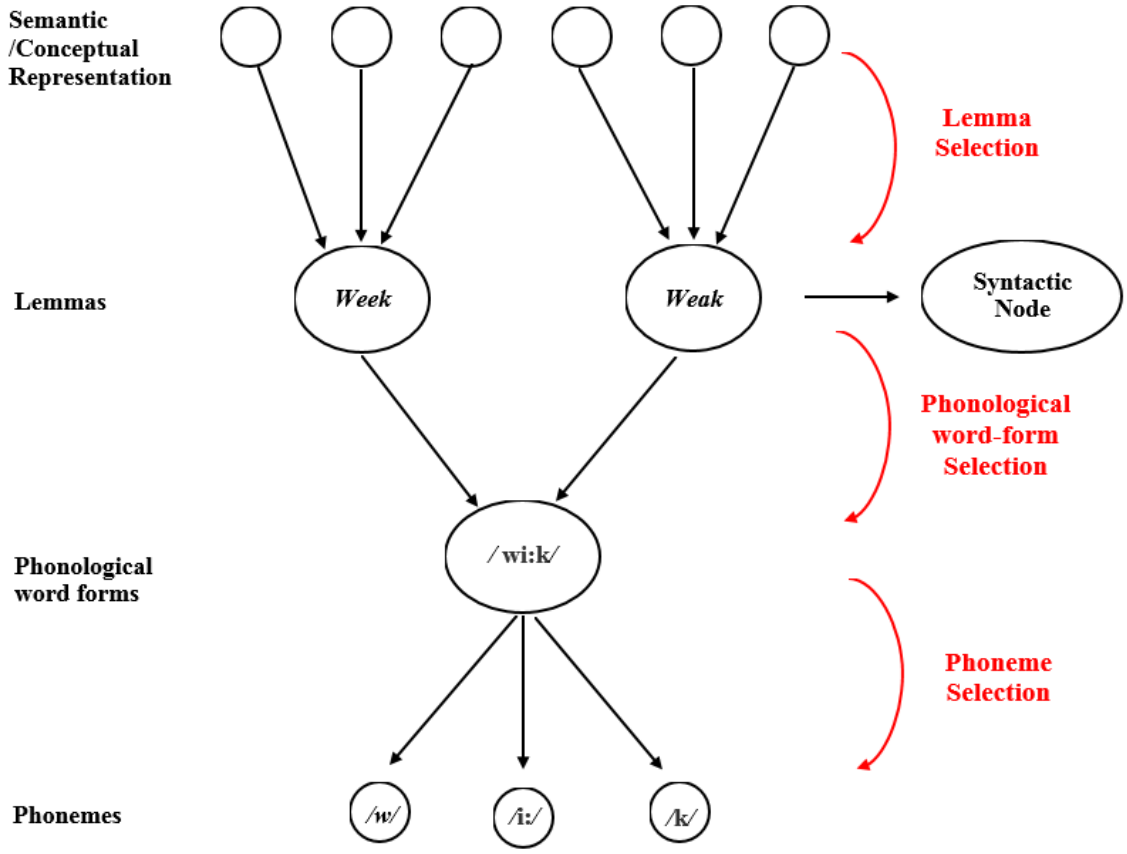



**Figure 2. 10** Outline representation of homophone frequency effect (based on Jescheniak, J. D., & Levelt, W. J. (1994). Word frequency effects in speech production: Retrieval of syntactic information and of phonological form. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*(4), 824. The word *week* shares the same phonological word-form with its homophone *weak*.

The first step to accessing the word-form is to complete the morphological encoding. For example, the morphemes of the verb *plant* in the past tense is <plant> and <ed>. This is followed by completing the metrical information which includes whether the word is monosyllabic or disyllabic, and where the stress is placed within the word. For example, the metrical information of <plant> is that it is a monosyllabic, unstressed morpheme. The last step in accessing the word-form is accessing the segmental information. The segmental information not only indicates the discrete consonants and vowels that are included in the morpheme <plant> (e.g. /p/, /l/, /ɑ:/, /n/, /t/) but also the order in which these need to be combined to constitute the phonological word (e.g. the order is /p/, /l/, /ɑ:/, /n/, /t/ but not /l/, /ɑ:/, /n/, /p/, /t/).

After encoding the phonological word-form, activation then passes down to the phoneme stage to accomplish the selection of phonemes. Phonemes are the individual speech sounds that make up the word; for example, /p/, /l/, /ɑ:/, /n/, /t/ are the phonemes that make up the word *plant*. The final stage is the execution stage, in which the sound wave that represents the word is articulated. If speech errors are involved in overt speech, people can detect the errors by self-monitoring. What's more, internal speech errors can also be detected by monitoring the phonological word-form before the onset of verbal production, halting before or during articulation to avoid speech error embarrassment (see Fig. 2.8).

In sum, the language production model of Levelt et al. (1999) starts from conceptual preparation, followed by the activation of semantic/concept representations. After completing lexical selection, activation then spreads to the lemma stage, and encodes the syntactic features of the selected lemma. The following stage is the phonological word-form stage. During this stage, the phonological shape of the word is retrieved. This is also the locus of the word frequency effect. Activation then carries to the next stage, the phoneme stage, and completes the phoneme encoding before articulating the target word. This model therefore proposes four main processing stages.

#### 2.3.3.2 The independent network model (Caramazza, 1997; Caramazza & Miozzo, 1997, 1998)

Most researchers favour the assumption that word production involves four stages: a conceptual stage, lexical stage, word-form stage and phoneme stage (Dell, 1990; Dell et al., 1997; Levelt et al., 1999). However, there is some disagreement about the number of stages involved in word production. Some researchers argue against the distinction between the lexical stage and the word-form stage where the lemma and whole sound shape are retrieved, respectively. These researchers instead suggest that the word and word-form are retrieved together in one stage, the lexeme stage (Caramazza, 1997; Caramazza & Miozzo, 1997, 1998). Therefore, as Caramazza (1997) proposes in the independent network model, there should be only three stages in language production: the conceptual stage, lexeme stage, and phoneme stage. Other crucial points proposed by Caramazza (1997) is that syntactic information is not included in lexical-semantic or word-form representations, but is represented independently. This contrasts with the model of

Levelt et al. (1999), who argue that syntactic features are part of the lemma properties (see Section 2.3.1).

Analysis of individual cases has demonstrated that brain-damaged subjects either have difficulty in accessing nouns while able to select verbs, or the other way around, suggesting a grammatical class dissociation between nouns and verbs (Berndt, Mitchum, Haendiges & Sandson, 1997a, 1997b; Caramazza & Hillis, 1991; Damasio & Tranel, 1993; Daniele, Giustolisi, Silveri, Colosimo & Gainotti, 1994; De Renzi & di Pellegrino, 1995; Hillis & Caramazza, 1995; Kremin & Basso, 1993; McCarthy &Warrington, 1985; Miceli, Silveri, Villa & Caramazza, 1984; Miceli, Silveri, Nocentini & Caramazza, 1988; Zingeser & Berndt, 1988). In addition, some brain-damaged subjects show difficulties in producing and/or receiving the same grammatical class words in different output modalities and/or input modalities (Hillis & Caramazza, 1991; Hillis & Caramazza, 1995; Rapp & Caramazza, 1997). For example, verbs can be accessed in oral production but not nouns, while nouns can be accessed in the writing modality but not verbs. In other words, both verbs and nouns can be accessed and lexical-sematic lexicon should be un-impaired. Based on the evidence that these deficits are restricted to specific modalities, Caramazza (1997) infers that it would be reasonable to assume that syntactic knowledge is represented independently from lexical-semantic knowledge. The possible account for the selective grammatical class deficits in different output modalities is that syntactic features are independent from lexical-semantic knowledge, but not associated with lemma. Therefore, as a consequence of this inference, the sequence of word production in this model involves the semantic/concept representation level, followed by the lexeme and syntactic levels (see Fig. 2.11).

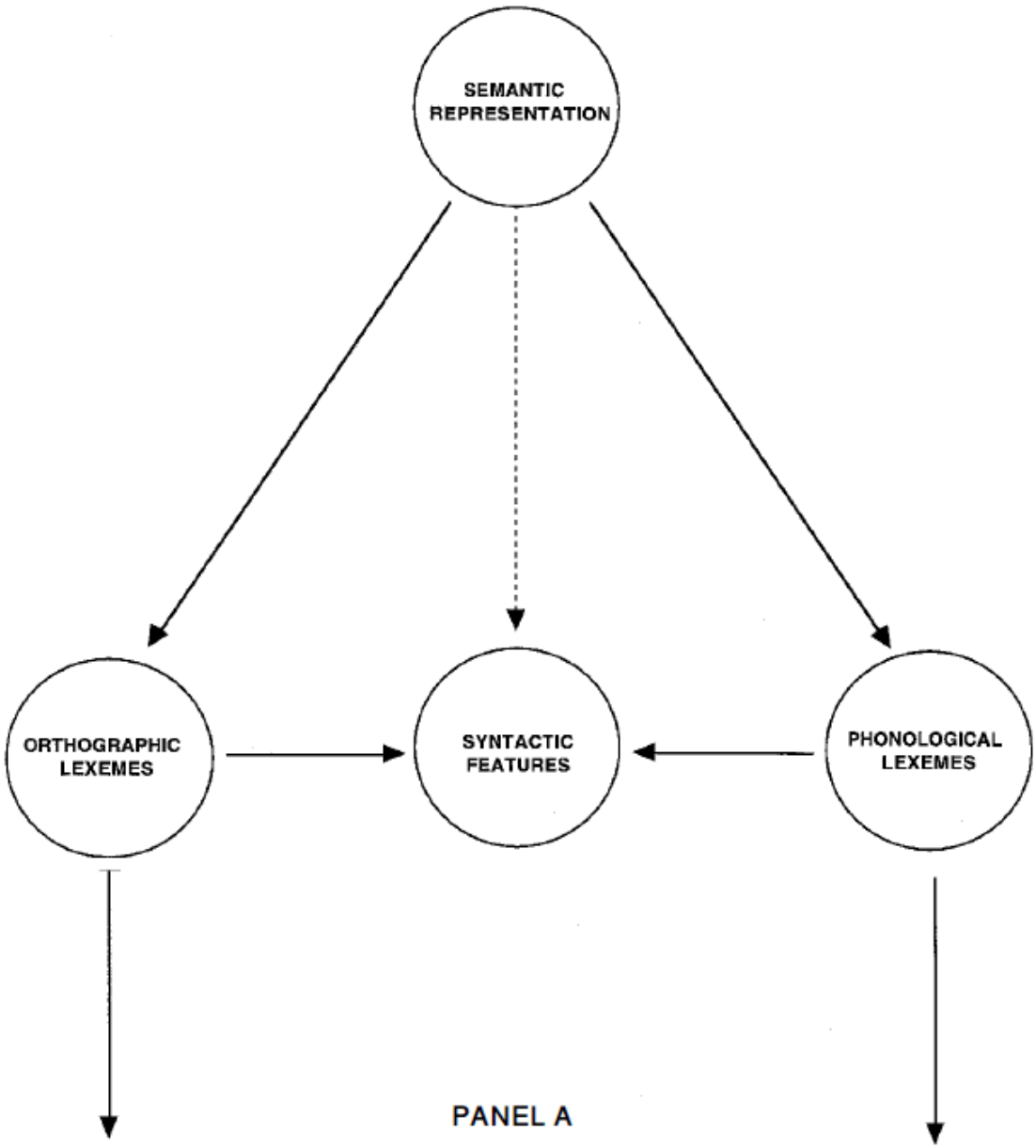



**Figure 2. 11** Outline of the independent network (IN) model which illustrates the relationship among semantic, syntactic, and lexeme representations. Adopted from Caramazza, A. (1997). How many levels of processing are there in lexical access? *Cognitive Neuropsychology, 14*(1), 177-208.

Like other models of lexical access, the independent network model starts from the semantic/concept representation level (see Fig. 2.12). Caramazza and Miozzo (1997) state that the semantic representations are decompositional, or componential (see Section 2.3.1), suggesting that multiple semantic features (BIRD, CAN FLY, NOCTURNAL, etc.) constitute the concept for an intended word (*owl*). The semantic representations send activation in parallel to the next level, the lexeme level, and activate a cohort of lexemes, which all share the semantic properties (or at least part of the semantic properties) of the

intended word, and other form information (namely, the structure of syllables). The activation is feed-forward only, but unlike the model proposed by Levelt and colleagues (1999), the activation takes place in a cascading way.

The most compelling evidence to support the cascade model comes from Peterson and Savoy (1998). In their study, participants were required to conduct a dual task by naming pictures and reading words. The SOA that separated the two tasks and words was manipulated. A facilitation effect of reading words was obtained when the word was either phonologically related to the target picture (e.g. the word *count* is phonologically related to the target picture COUCH) or phonologically related to synonyms of the target picture (e.g. SOFA is a synonym of COUCH; the word *soda* is phonologically related to SOFA, the synonym of the picture COUCH) at short SOA, suggesting that phonological representations of the unintended word SOFA are activated, and therefore, facilitate the word reading latency. However, in the long SOA condition, a facilitation effect is only found when the word is phonologically related to the target picture, suggesting that unintended words' phonological nodes are not activated after fulfilling the lemma selection. Thus, reading word latency should not be facilitated if the word is phonologically related to the synonym of the picture.

Importantly, Caramazza (1997) also proposed the crucial assumption that there is a positive correlation between the amounts of activation that the lexeme will receive and the number of features that the lexeme shares with the intended word, so that the more features the lexeme shares with the intended word, the more activation it will receive. Each feature receives an equally divided amount of activation propagated from the semantic/concept representation level. The threshold that lexemes need to reach requires the full amount of activation being passed down from the semantic/concept representation level. For example, suppose the intended word *plant* has five features, and each feature receives 1/5 the amount of activation. The word *flower* shares 3 features with the word *plant*, and receives 3/5 the amount of activation. Since the full amount of activation is needed to reach the threshold, only the word *plant*, but not *flower*, can pass the threshold.

In order to offer an account of different production modalities, two modalities of lexemes are introduced in this language production model: phonological lexemes (P-lexemes) for oral production and orthographic lexemes (O-lexemes) for written production. Both these

types of lexemes receive activation from the previous level and fulfil the independent selection of the specified lexeme representation. Meanwhile, independently represented syntactic features (e.g. what tense the word is; whether it is plural or singular, a verb, adjective or noun; or masculine or feminine) also receive weak activation from the selected semantic representation which can prime these features. However, only the activation that lexemes spread to syntactic features allows the selection of the intended syntactic features.

In sum, the sequence of oral word production in the independent network model involves two stages: after the intention to produce a word, semantic/concept representations spread activation to syntactic features, P-lexemes and O-lexemes. After the selection of the appropriate lexemes (lexeme selection) with the intended modality, and syntactic features, the activation spreads to the phoneme level and the appropriate phonemes are selected (phoneme selection). Therefore, there are only two selection stages in this word production model.

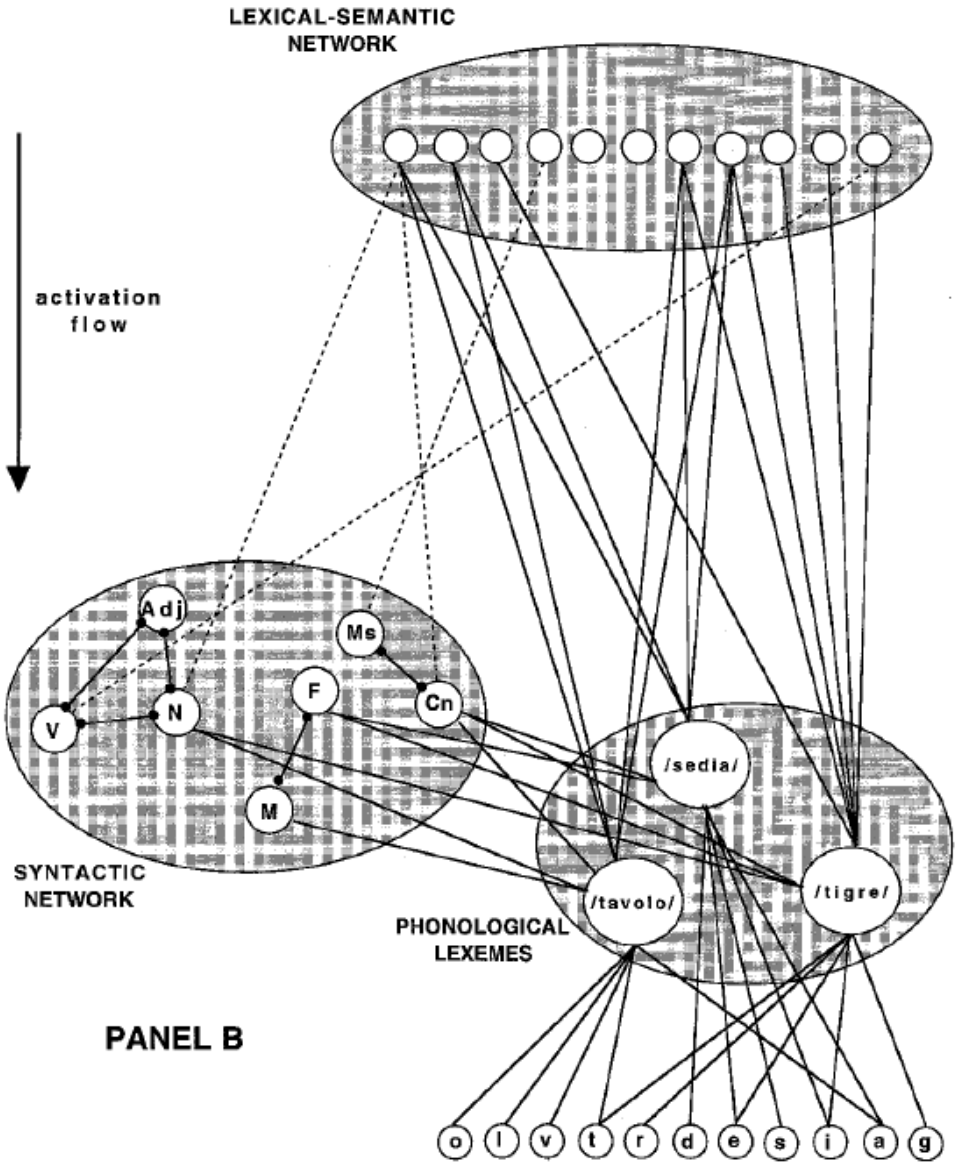



**Figure 2. 12** Independent network (IN) model of oral language production. The activation spreads from semantic to lexeme and syntactic networks, and then passes down to segmental information. Dotted lines represent weak activation. N=noun; V=verb; Adj=adjective; M=masculine; F=feminine; CN=count noun; Ms=mass noun. Adopted from Caramazza, A. (1997). How many levels of processing are there in lexical access? *Cognitive Neuropsychology*, *14*(1), 177-208.

The above word production theories for monolinguals are well established on a large number of empirical research results as well as the observation data of speech errors. However, with an increasing population of bilinguals, it is necessary to explore the processing procedure of language production of bilinguals, with a particular emphasis on determining whether there is a difference between monolinguals and bilinguals. The following section will further discuss the language production of bilinguals.

### 2.3.4 An introduction to bilingual lexical access

#### 2.3.4.1 Introduction to key issues

When bilinguals intend to speak, for example, to name the picture MONKEY, they not only need to select the appropriate name of the picture, as monolinguals do, but also need to decide on the language, since they can speak more than one language. The basic word production architecture and processing procedures of bilinguals should be similar to monolinguals, with the exception of one central question that requires an answer: How are bilinguals able to select the intended word (e.g. *monkey*) in the target language rather than its translation equivalent in another language (e.g. *houzi*, the equivalent of *monkey* in Chinese)? This problem has been termed the "hard problem" (for details, see Finkbeiner, Gollan & Caramazza, 2006). Heated debates have raged about bilingual lexical access, mainly centred on this problem, and several models have been provided to account for the "hard problem" (e.g. Costa, Miozzo & Caramazza, 1999; Green, 1986; 1998; La Heij, 2005).

Earlier proposals favoured the idea that only the target language lexicon will be activated when bilinguals intend to speak (MacNamara & Kushnir, 1972). However, the assumption that two languages are activated in parallel during the lexical stage in bilinguals regardless of the target language has been widely accepted, based on a large amount of research data (Costa et al., 1999; Costa & Caramazza, 1999; Cutting & Ferreira, 1999; De Bot, 1992; Green, 1998; Hermans et al., 1998; Kroll & Stewart, 1994; La Heij, 2005; Poulisse, 1997, 1999). According to this assumption, it is reasonable to infer that each (common) semantic representation is shared by two lexical nodes, and each lexical node is stored in each language lexicon (De Bot, 2000; De Groot & Nas, 1991; Grosjean, 2001; Kroll & Stewart, 1994). However, it is still possible that some conceptual representations are language/culture dependent, as some researchers have indicated (Van Hell & De Groot, 1998a; Jared, Poh & Paivio, 2013).

It is claimed that when bilinguals intend to produce a word, both semantically related lexical representations in two languages are activated, even followed by a sentence (Starreveld, De Groot, Rossmark & Van Hell, 2014). Therefore, the question arises of how bilinguals separate the lexicons of two languages and prevent intrusions from the language

that is not in use. The controversy of bilingual language production centres mainly around whether lexical selection is language-specific (the selection of a word takes place within the target language – Costa et al., 1999; Costa and Caramazza, 1999; Finkbeiner, Almeida, Janssen & Caramazza, 2006), or is language-non-specific (selection takes places between two languages – Dijkstra & Van Heuven, 1998; Green, 1998; Kroll, Bobb, Misra, & Guo, 2008; Meuter & Allport, 1999; Grosjean, 2001). To further explore this controversy, the empirical data from research on bilingual lexical access will first be briefly summarised in Section 2.3.4.2. Then the bilingual language production models that are in line with different assumptions of lexical selection will be discussed in Section 2.3.5.

### 2.3.4.2 Brief summary of the behavioural results of bilingual lexical access

The same methodologies have been adopted to explore the lexical access of both monolinguals and bilinguals. Beside the methods of speech errors (Brosseau-Lapré & Rvachew, 2014; Poulisse, 1999), the TOT phenomenon (Gollan, Ferreira, Cera & Flett, 2014), and the picture-word interference paradigm (Costa and Caramazza, 1999), language switch tasks have also been adopted to further explore whether lexical access is language-specific or non-specific. The critical results obtained from using picture-word interference paradigms and language switch tasks will be discussed in the following paragraphs.[3]

Evidence from picture-word interference tasks indicates that when the distractor is the name of the picture, either in the target language or non-target language, the response latency is faster compared to an unrelated control distractor in both the blocked condition (the target language is consistent during the block) and the mixed condition (the target language is decided by the cue during the block). This kind of facilitation is called the **identity effect**. The facilitation of a target-language identical distractor occurs at negative, 0, and positive SOA (in other words, the represented distractor appears earlier, simultaneously, and later than the picture, relatively) (Starreveld & La Heij, 1996), while the identity effect occurs only at negative SOA when the distractor is an identical word in a non-target language. Moreover, the identity effect is larger in the presence of target language pairs than in the presence of different language pairs.

[3] Studies of word translation are not included in this section, due to their limited relevance to the thesis. However, for research on word translation, see, for example, Kroll and Stewart (1994), La Heij, Hooglander, Kerling, and Van Der Velden (1996).

The reaction time is slower under non-cognate semantically related distractor conditions than under unrelated control conditions. The magnitude of the **semantic interference effect** is typically the same whichever language the distractor is presented in, consistent with the assumption that activation spreads to translation equivalent lexical nodes in two languages equally in proficient bilinguals (Costa & Caramazza, 1999; Costa, Santesteban, & Ivanova, 2006). However, when the distractor word is a cognate word, which is orthographically and phonologically similar to its translation in another language, for example, *dag* [Dutch, 'day'] is the cognate word for *day* [English, 'day'] while *afval* [Dutch, 'waste'] is the non-cognate word for *waste* [English, 'waste'], a **cognate effect** is obtained. The reaction time to cognate picture names is shorter than to non-cognate names, regardless of what the bilingual's dominant language is (Costa, Caramazza, & Sebastian-Galles, 2000).

The phonological effect is known to facilitate the reaction of picture naming in the same language, as mentioned in Section 2.3.2. Interestingly, however, when the distractor is phonologically related to the translation equivalent of the target word, an interference effect is obtained, rather than a facilitation effect. This interesting result has been found by Hermans, Bongaerts, De Bot, & Schreuder (1998), who showed that when the picture *mountain* [English, 'mountain'][4] was presented and required Dutch-English bilinguals to respond in English, the distractor word *berm* [Dutch, 'verge'], which is phonologically related to the Dutch word *berg* [Dutch, 'mountain'], slowed the picture naming more than the unrelated control distractor *kaars* [Dutch, 'candle'].

For language switching tasks, the commonly adopted paradigm is the cued language switching paradigm which requires participants to name the presented stimulus in the language according to the cue. The difference between a repeated language trial and a switched language trial is called a "switch cost". A large number of research results have shown that the reaction time to name a picture in the same language as the preceding one is

[4] In the phrase "*mountain* [English, 'mountain']", the word in the square brackets represents the language property of the word preceding the square brackets, and its English translation. Thus, the word *mountain* is an English word and its English translation is "mountain".

normally faster than after a switch into another language (e.g. Christoffels et al., 2007; Guo, Liu, Chen & Li, 2013; Jackson et al., 2001; Linck, Schwieter & Sunderman, 2012; Meuter & Allport, 1999). However, symmetrical switching costs are obtained in some studies with proficient bilinguals, showing the reaction time of switching the language is as fast as repeating the language (Christoffels, Firk & Schiller, 2007; Costa & Santesteban, 2004) even when comparing the strong second language and weaker third language (Costa, Santesteban & Ivanova, 2006).

### 2.3.5 Theories of bilingual lexical access

To offer an account of the critical research results outlined in Section 2.3.4.2, two influential bilingual lexical access models have been proposed, namely **the language-specific selection model** (Costa, Miozzo & Caramazza, 1999) and **the inhibitory control (IC) model** (Green, 1998). These two models are in line with the assumption that language selection is specific and non-specific, respectively, and the basic concepts are discussed in the following sections, in turn.

#### 2.3.5.1 The language-specific selection model (Costa, Miozzo & Caramazza, 1999; Costa & Caramazza, 1999)

Since lexical representations are activated in the lexicon of both languages during language production, the language-specific selection model suggests that nodes in both languages will be activated to equal degrees. However, bilinguals will only select the target word within the target language lexicon, whether the target language is the first language or second language, and even when the two languages are highly similar to each other (Costa, Miozzo & Caramazza, 1999; Costa & Caramazza, 1999; Costa, Colomé & Caramazza, 2000).

This language-specific selection assumption is strongly supported by the results of the identity effect. If lexical selection is between two languages, then the translation equivalent word in the target language should be the most competitive distractor since it can receive activation from both picture and distractor word. Contrary to this expectation, robust

results have shown that translation equivalent distractors actually facilitate the response latency rather than interfere with it. What's more, the selection/control of language, therefore, should happen at or precede the lexical stage (for review, see Declerck & Philipp, 2015), otherwise the translation equivalent lexical node will compete with the target node.

However, as mentioned above, the identity effect is different depending on whether the language of the distractor word is in the target language or in the non-target language. If nodes in two language lexicons receive equal activation, as Costa and colleagues (1999) indicate, the identity effect should be the same, whether the distractor is presented in the target or non-target language. Costa and colleagues (1999) attribute the difference between target and non-target language identity effects to phonological facilitation. In this interpretation, target language identical distractors activate the intended word's orthographic representation and spreads activation to both the lexical and sub-lexical phonological stage, while non-target language identical distractors can only send activation to the lexical stage. Therefore, the identity effect is larger for same-language pairs than different language pairs, regardless of the SOA condition.

Furthermore, another critical hypothesis has been proposed, namely that a word distractor can activate both itself and its translation equivalent word. This feature is called "automatic translation" (Costa, Miozzo & Caramazza, 1999). Automatic translation offers an explanation why different-language semantically related distractors can result in an equal size of semantic interference effect as same-language distractors, since selection is conducted within the target language.

Similar to the independent network model (Caramazza, 1997; Caramazza & Miozzo, 1997, 1998), the bilingual lexical access model also favours a cascaded notion of activation, with the supporting results of cognate effect. In accordance with the assumption that phonological representations of non-selected lexical nodes in the non-target language lexicon are also activated (Colomé, 2001; Costa, Caramazza, & Sebastian-Galles, 2000; Costa, Roelstraete, & Hartsuiker, 2006), the phonological nodes of cognate words should be retrieved faster compared to non-cognate words, since phonological segment retrieval depends on the degree of activation. Research results are consistent with the assumption. Furthermore, the potential explanation that there are intrinsic differences between cognate

and non-cognate pictures among monolinguals has been ruled out (Costa, Caramazza, & Sebastian-Galles, 2000).

The phonological interference effect shown by Hermans and colleagues (1998) seems to contradict the language-specific lexical selection hypothesis in the first instance, since it demonstrates that lexical selection occurs between the two languages. However, Costa and colleagues (2000) argue that the interference effect might be due to the competition among phonological representations but not due to lexical competition between two languages. This assumption has been supported by the evidence showing reaction time is slower when presenting a phoneme that is in the name of the picture in the other language during a decision task (Colomé, 2001). This suggests that the activated translation-equivalent lexical node in the non-target language spreads activation to its phonological representation, and as a result, this phoneme is harder to reject compared to an unrelated target phoneme.

In offering an explanation to the asymmetrical switching costs widely obtained in the language switching task, Costa and Santesteban (2004) provided a comparison between the performance of proficient bilinguals and second language (L2) learners in that task. Their results show asymmetrical switching costs are presented among L2 learners. For proficient bilinguals, however, symmetrical switching costs are obtained when testing one dominant language with another dominant language or even with another weaker language. Costa and Santesteban (2004) suggest that some mechanisms of proficient bilinguals and L2 learners might be different, as well as processing procedures of their lexical access.

In sum, the language-specific lexical selection model favours the assumption that activation flows in a cascaded way and corresponding segments receive the activation before the completion of lexical selection. Two existing language lexicons are activated during speech, and each language lexicon receives equal activation from the previous conceptual stage. Activated lexical representations from two languages pass the activation down to their phonological segments. However, lexical selection proceeds only within the target language. In addition, two translation equivalent words can be activated automatically by a word regardless of the language.

#### 2.3.5.2 Inhibitory control (IC) model (Green, 1986, 1998)

The inhibitory control (IC) model offers an alternative account of bilingual language production. It shares the assumption that multiple levels are involved, and that two languages' lexicons receive activation during language production (i.e. the competition assumption). In contrast with the language-specific selection model (Costa, Miozzo & Caramazza, 1999; Costa & Caramazza, 1999), however, the IC model argues that lexical selection is between languages, and the dominant language (L1) receives a higher level of activation than the non-dominant language (L2).

One of the basic assumptions of Green (1998) is that language processing and behavioural actions have much in common, as "language is a form of communicative action" (p. 68). Both non-verbal action and verbal action are subject to voluntary control, require a goal and cause a precise action. In accordance with the action model proposed by Norman and Shallice (1986), Green indicates that multiple levels of control and a language task schema are involved in the IC model during language production. A language task schema, as defined by Green (1998), refers to "mental devices or networks that individuals may construct or adapt on the spot in order to achieve a specific task and not simply to structures in long-term memory" (p. 69). A schema can either be retrieved from long-term memory (e.g. when communicating in one language) or from the instruction of task (e.g. when required to press "Red" when hearing a high-pitched tone and to press "Green" when hearing a low-pitched tone).

To achieve the goal (G) of communication, the **supervisory attentional system (SAS)** mediates the activation level of the control language task schemas competing for output. For example, when interpreters intend to conduct SI, the translation schema competes with the repeating schema to control the output from **lexico-semantic systems (**Green, 1998**)**. SAS alters the activation degree of both these schemas to achieve the selection of the intended translation schema. If the goal has changed, then the originally selected schema will be inhibited by SAS. If, however, the goal is achieved, then the selected schema will be supressed by the schema itself.

Evidence from studies of individual aphasics (Paradis, Goldblum & Abidi, 1982) favours the assumption that language task schemas are separable and competitive. Results showed

that though some bilingual aphasics can understand two languages, they can spontaneously produce only one language on one day, but not on the subsequent day. It means on Day 1 they can only speak in their L1, while in Day 2 they can only communicate in L2. However, on Day 3, they can use their L1 again but not L2. More interestingly, subjects are able to correctly translate into the language that they cannot spontaneously produce without hesitation. However, they demonstrate very poor performance when translating into the language that is available for spontaneous use. Thus, the pattern of findings indicate that spontaneous language production is functionally different from translation production, suggesting that the translation task schema and language production schema are separate.

The Stroop effect is also consistent with the schema separation and competition assumptions. In the Stroop task, when the presented string, which is the name of the colour, is not identical to the name of the string's colour (e.g. RED), the reaction time of the response to the colour is postponed compared to the condition where the string is a nonword (e.g. XXXX). The interference of the response time is known as the Stroop effect. Green (1998) suggests that part of the Stroop effect may be due to the competition between the reading task schema, which is triggered automatically by the string stimulus, and the colour name schema, which is elicited by the instructions.

After the selection of the intended language task schema, the schema exerts its output control by mediating activation according to the tag specification. **A language tag** is postulated to be associated with a lemma (Green, 1986, 1993; see also Poulisse & Bongaerts, 1994) and specifies whether the lemma belongs to L1 or L2. The locus of language selection is, accordingly, at the lemma stage. The lemma representations with the intended language tag (e.g. L2) receive activation while lemmas with the incorrect language tag (e.g. L1) are suppressed (see Abutalebi, & Green, 2007 for neuroimaging discussion). The amount of suppression generated by SAS is correlated with the activation level of the lemmas with the incorrect language tags, suggesting that if those lemma representations which are associated with incorrect language tags are strongly activated (e.g. L1, the dominant language of unbalanced bilinguals receives higher activation than L2), then the suppression that they will receive is also greater to achieve the goal of selecting the correct language. That is, when intending to reactivate the previously inhibited language tag, the more supressed language tag (e.g. dominant language) takes

longer to reactivate than the less inhibited language tag (e.g. non-dominant language). This assumption is consistent with the results of asymmetrical switching costs in language switching tasks.

Stemming from the IC model, it has been proposed that bilingualism confers processing advantages in non-linguistic tasks (Bialystok, 2001). The inference of this assumption is simple: the dominant-general executive function system, the system which resolves both linguistic and non-linguistic conflict tasks, has been highly practiced by bilinguals, due to inhibiting one language while activating another language. Therefore, bilinguals should outperform their monolingual counterparts as a consequence (but see Antón, Duñabeitia, Estévez, Hernández, Castillo, Fuentes, & Carreiras, 2014; Donnelly, 2016; Hilchey & Klein, 2011; Kousaie & Phillips, 2012; Mor, Yitzhaki-Amsalem, & Prior, 2014; Paap & Greenberg, 2013). This proposal has been widely explored on a large scale and across the lifespan in recent years, and a large number of empirical results which favour bilingual advantages offer further indirect support for inhibitory control (Bialystok, Craik, Klein & Vishwanathan, 2004; Bialystok, Craik, & Luk, 2008, 2012; Engel de Abreu, Cruz-Santos, Tourinho, Martin & Bialystok, 2012; Gold, Kim, Johnson, Kryscio, and Smith, 2013; Luk, Bialystok, Craik, & Grady, 2011; Martin-Rhee & Bialystok, 2008; Pelham & Abrams, 2014; Poarch & van Hell, 2012; Schroeder & Marian, 2012; Tao, Marzecová, Taft, Asanowicz & Wodniecka, 2011; Yang, Yang, & Lust, 2011; see Bialystok, 2010 for review).

In sum, the IC model suggests that both languages receive activation during language production, and the amounts of activation that each language receives have a positive correlation with its degree of proficiency. SAS modulates the activation of lemmas in accordance with their language tag and suppresses the lemmas that are associated with an incorrect language tag.

### 2.3.6 Summary of word production models

Clear insights can be gained after reviewing the research on lexical access in both monolinguals and bilinguals. Multiple stages are involved even when producing a single word. The starting point of word production is the conceptual stage. A group of conceptually related representations are activated with the target word. Activation then

passes to the following lemma stage and activates the connected lemma representations. The meaning and syntactic features of the activated lemmas can be encoded at this stage. The whole sound segment is retrieved in the next stage, the phonological-word-form stage, where high frequency words are more easily retrieved than low frequency words. Each individual sound or phoneme of the lemmas, however, is encoded in the phoneme stage, the last processing stage of word production. Before the commencement of articulation, the selection of the appropriate lemma, phonological word-form, as well as phoneme is necessarily required. If the incorrect selection is made, then speech errors will be produced as a consequence.

People who can speak more than one language also need to guarantee the appropriate word is selected within the intended language. If the lexical selection is within the target language, as the language-specific selection model (Costa, Miozzo & Caramazza, 1999; Costa & Caramazza, 1999) suggests, highly proficient bilinguals are also required to fulfil the selection of languages in addition to conducting the word production procedures as monolinguals do. Therefore, highly proficient bilinguals should be slower in word production than monolinguals even in their dominant and first language. This interference is consistent with empirical research results (Ivanova & Costa, 2008). However, if lexical selection is between languages, in accordance with the indication of the IC model (Green, 1986, 1998), lexical representations in two languages' lexicons are activated and compete for selection. If bilinguals intend to produce the weaker language (e.g. L2), then stronger inhibition is required upon the dominant language (e.g. L1) to achieve the goal of correctly producing the non-dominant language. Therefore, a longer response latency is needed for bilinguals to produce a word in the weaker language than in the dominant language, and it surely takes more time for them to produce a word than for monolinguals, even in their dominant language.

Professional simultaneous interpreters should no doubt be classified as proficient bilinguals. If the lexical selection of professional simultaneous interpreters is within the target language, as indicated by the language-specific selection model, then both interpreter and bi-/multilingual groups should take more time to produce a word compared to monolinguals. However, if simultaneous interpreters adopt the lexical access procedures as the inhibitory control model suggests, then there is a possible hypothesis that these special groups might have advantages in inhibiting the non-target language thanks to their

exposure to the conditions of continuously receiving the SL while producing into the TL. Consequently, their response latency may be as fast as monolinguals, and shorter than their proficient bilingual counterparts. It is also possible that the selection of lemma and phonological-word-form should be easier for simultaneous interpreters than for bilinguals since, if simultaneous interpreters have advantages in inhibiting the non-target language, the lexical competitors for selection may possibly only be within the target language lexicon, but not across languages like bilinguals.

Interestingly, independent, unrelated non-linguistic tasks have been demonstrated to interfere with the performance of word production. Lemma selection, phonological word-form selection (Ferreira & Pashler, 2002), as well as phoneme selection (Cook & Meyer, 2008) have all been shown to be subject to the central processing bottleneck, the stage where only one task can be conducted each time. Further evidence comes from word-picture interference tasks (Kleinman, 2013), suggesting that lexical selection is subject to the central processing bottleneck. Some research has even shown that the accuracy of recall hinges on, and the reaction time of recall can be postponed by, a non-linguistic task (Carrier & Pashler, 1995; Rohrer et al., 1998; Rohrer & Pashler, 2003), demonstrating that the bottleneck has an impact on memory retention as well as memory recall.

As mentioned in Section 2.2, during SI, comprehension, memorising, and production take place at the same time. If production has an impact on short-term memory and postpones the reaction time for recall, then parts of the message which needs to be delivered may be reduced or omitted due to memory damage and/or time pressure. Consequently, the quality of the interpreting may suffer. However, the interpretation of professional interpreters normally covers the important information that the speaker intends to deliver, not to mention that it is produced in real time. Thus, the high quality performance of simultaneous interpreters begs the question: Are simultaneous interpreters also subject to the central processing bottleneck during language production? Before moving on to explore this question, the concept of the central processing bottleneck, together with the dual-task methodology that has been used in the field of research focusing on divided attention will be discussed in more detail in the next section.

## 2.4 Introduction to dual task performance

In daily life, people commonly perform more than one task at the same time without much difficulty. They can drive while talking to a passenger, or reply to an email while chatting with their friends. However, there is general agreement that the difficulty in performing two tasks may increase when either or both of the tasks are complicated (such as answering the questions in a television quiz while answering a business email). Based on these observations, some interesting questions have been raised: How many things can people do simultaneously? What happens when people are conducting more than one task at the same time? How is performance affected when two tasks are carried out at the same time?

Consciously dividing and allocating attention to performing different but concurrent tasks is termed "divided attention". Research on divided attention has been conducted for a long time to explore the limitations of human capacity as well as the functional architecture of the human brain (Pashler, 1993). Research results show that there are fewer things that people can do concurrently than they think. The performance of at least one task usually suffers when conducting two concurrent or closely sequentially presented simple choice tasks that require speeded responses. Either/both the accuracy of the responses is damaged or/and the reaction time is delayed. This section focuses on whether people can perform two tasks at the same time without any interference. It is organised in two sub-sections. Section 2.4.1 provides a brief overview of research on dual task performance, and the interference caused by conducting two tasks at the same time. This is termed the psychological refractory period effect (PRP effect). Section 2.4.2 introduces the two most popular divided-attention models that account for the PRP effect, and discusses these two models in some detail.

### 2.4.1 Dual task performance and the Psychological Refractory Period (PRP) effect

As pointed out above, people often perform two tasks simultaneously without any interference (e.g. write while listing to music) or do two things at the same time in quite an efficient way (e.g. chatting while reading books). People can normally perform two tasks well when they have plenty of time, since they can begin the second task after finishing the first task. The question is: What happens when conducting two simple tasks at the same time, with time pressure?

This question has been widely explored in the field of psychology, adopting a typical dual-task paradigm, shown in Figure 2.13. In each choice reaction time task, the subject is required to make a specific response to the present stimulus (e.g. pressing button "B", "N", and "M" for stimulus letter "A", "B" and "C", respectively; saying "High" or "Low" for a high or low pitch tone, respectively). Each choice reaction time trial involves three basic processing stages: the **perceptual stage, response selection stage, and response execution stage**. The **perceptual stage** begins by commencing perceptual processing immediately after each stimulus arrives. After finishing encoding the stimulus, the central mechanism will be occupied and begins to select the appropriate response. This stage is called the **response selection stage**. When the response selection is complete, the motor action of responding to the stimulus will be conducted, which is known as the **response execution stage**.

When the stimuli from two discrete tasks (Task 1 and Task 2) are presented in rapid succession, the response time to the second stimulus is usually delayed as SOA is reduced, whereas the first response is fairly independent of manipulating the SOA (shown in Fig. 2.14). In other words, when decreasing the interval between the onsets of two different tasks, the reaction time to the second stimulus increases, but not the reaction time of the first task. The increase in reaction time of the second task is maximally prolonged when the two tasks are presented simultaneously (the interval between the two tasks is zero), compared to the situation when two tasks are displayed separately with a long interval between them. This phenomenon was first demonstrated by Telford (1931), who termed it the **psychological refractory period** (PRP), in accordance with the refractory period of neurons. This phenomenon essentially means that when two different tasks are presented closely in time, the performance of the second task is impacted, suggesting that people cannot do two things at the same time without any interference.

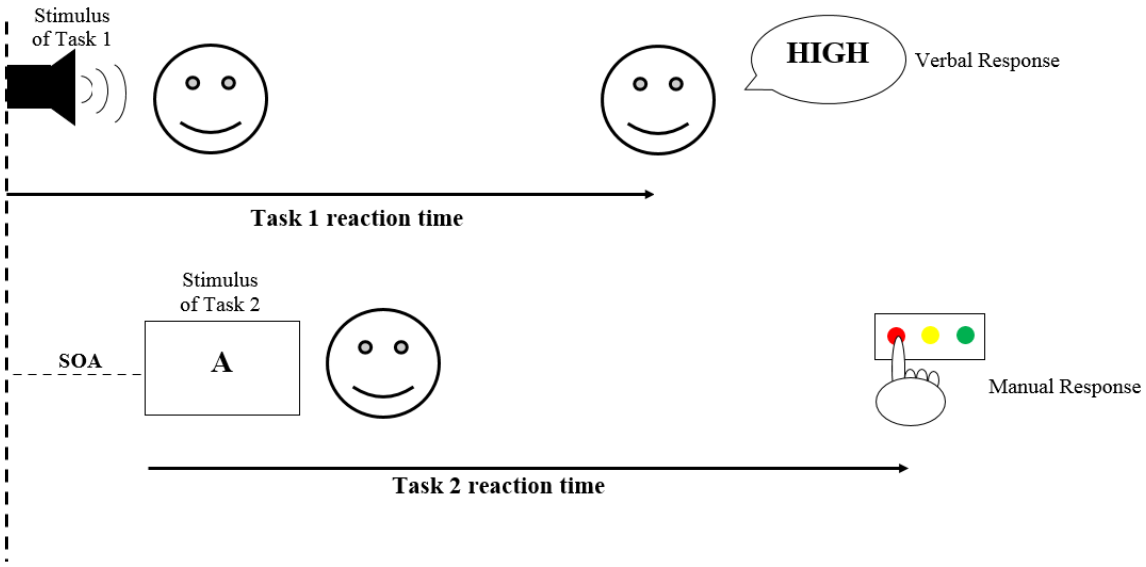



**Figure 2. 13** Typical dual task paradigm: Two choice tasks are separated by varying SOA. Task 1 plays a pitch tone and requires a verbal participant response, by saying "HIGH" or "LOW" according to the high pitch tone or low pitch tone, respectively. Task 2 presents a capital letter "A", "B" or "C" and asks the participant to press a red, yellow, or green button respectively.

The PRP effect has been tested in a great variety of different non-linguistic choice reaction time tasks in the past with very diverse speeded response modalities (for review, see Pashler, 1994) including manual, by pressing buttons (Lee & Chabris, 2013; Liepelt, Prinz, 2011; Pashler, 1984; Strobach, et al., 2015) or drawing lines (Vince, 1948); vocal (e.g. saying "High" or "Low"; see Pashler, 1990; Pashler & Christian, 1994; Levy, Pashler & Boer, 2006); eye-movement (Pashler, Carrier & Hoffman, 1993), and even foot responses (e.g. stepping on a car brake pedal; see Levy, Pashler & Boer, 2006; Osman & Moore, 1993). Evidence shows that when two tasks require the same response modality (e.g. both tasks require a manual-press response), the interference is just a little larger than when each task requires different response modalities (e.g. one requires a verbal response and one requires a manual response) when the order of the stimuli from the two tasks is certain (Pashler, 1990). In this scenario, differences in response modality for the two tasks do not significantly diminish interference. The possibility that dual task interference is caused by the unpredictable presentation of the second task is ruled out by setting a constant SOA during the whole block (Bertelson, 1967; but see Koch, 1995, cited in Navon & Miller, 2002). These results indicate that the PRP effect is robust regardless of the response

modality of two tasks, and this phenomenon is evident even in simple reaction time tasks (Telford, 1931).[5]

Different from the typical dual task which presents one stimulus and requires one independent response in each of the two tasks, dichotic listening tasks involve only a single task which presents multi-stimuli and requires one response. The dichotic listening task has also been widely adopted to explore divided attention since the 1950s (Broadbent, 1958; Deutsch & Deutsch, 1963; Treisman, 1964), in parallel with the dual-task paradigm. In this kind of task, three different numbers are played to participants' one ear over a headset, while another three numbers are simultaneously presented to the other ear. For example, the numbers "1, 9, 6" may be played to a participant's right ear while the digits "5, 2, 7" are presented to their left ear. The participant is required to recall the numbers from one of the ears, based on an instruction. Evidence has shown that if two tasks have the same stimulus modality (e.g. both tasks present visual or auditory stimuli), then the interference is comparably larger than the situation when two tasks display different stimulus modalities (e.g. one task presents a visual stimulus and the other presents an auditory stimulus) (Treisman & Davies, 1973; Wickens, 1980). The more favoured explanation accounting for the occurrence of interference in dichotic listening tasks is suggested to be the reason of filtering unintended information at perceptual stage with strong evidence provided by Treisman and Riley (1969).

Impaired performance in the multi-stimuli task, like the dichotic listening task, is suggested as the result of exceeding the limited processing capacity at the perceptual stage. However, is the impacted performance of the dual task caused by the same reason as the dichotic listening task, since they are all conducted in a divided attention condition? Or does a different type of interference occur in accordance to different types of tasks? Several possibilities have been provided to account for dual task interference in the past 60 years. In the following section, two of the most influential explanations accounting for dual task

[5] Interference effects may not be particularly salient to people. Completing a simple choice task, such as that normally conducted in an experimental setting (for example, pressing different buttons in response to different letters), usually takes people less than a second. The interference caused by conducting a dual task increases in the same speed ratio to the SOA. If the interval between two tasks decreases by 1 millisecond, then the reaction time to the second stimulus increases by only 1 millisecond. Therefore, people may not realise the interference cost by conducting dual tasks unless the tasks are either a) highly similar to each other (share the same response modality), or b) demand considerable intellectual effort to resolve.

interference will be summarised and discussed: the bottleneck model and the capacity sharing model.

### 2.4.2 Models accounting for dual task interference

#### 2.4.2.1 The bottleneck model

One of the most influential theories to account for the PRP effect is the bottleneck model. The model favours the assumption that dual task interference is the result of being incapable of processing two specific mental operations at the same time. Parallel processing cannot take place in the dual task condition because a single mental mechanism is required by a processing stage or bottleneck stage (e.g. response selection stage) in both tasks. When the two tasks overlap with each other, the bottleneck stage (e.g. response selection stage) in the second task (T2) has to wait until the same bottleneck stage in the first task (T1) is completed, regardless of whether the previous stage of T2 is complete or not (See Fig. 2.14), while the non-bottleneck stage(s) can be conducted with other stage(s) in parallel (e.g. perceptual processing can be overlapped with the response selection stage; Levy & Pashler, 1995). In other words, the bottleneck stage (e.g. response selection stage) is postponed but the previous stage can be processed. Therefore, the interference occurs as a result of the postponement of the second task in the overlapping task condition.

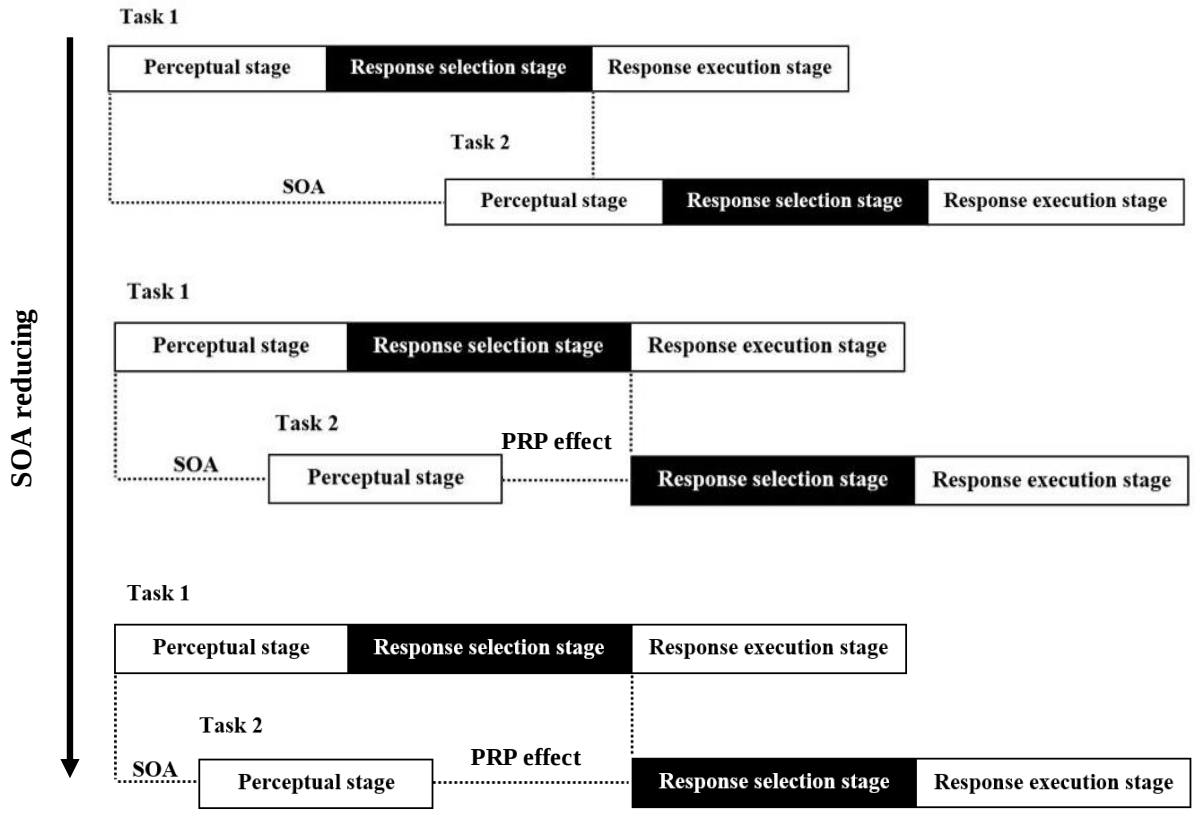



**Figure 2. 14** A central bottleneck model (for review, see Pashler, 1994): Response selection stage of T2 cannot begin until the corresponding stage of T1 is complete while other stages can operate in parallel.

The controversy around the bottleneck model is mainly centred on the processing stage where the bottleneck is located. Some favour the assumption that the bottleneck or central processing bottleneck is located at the **perceptual stage** (Broadbent, 1958; Keele, 1973; but see Ruthruff, Miller & Lachmann, 1994; Stroop, 1935), suggesting that the perceptual processing of the second task cannot begin until the perception of the first task is finished. Others argue that the bottleneck is involved at the **response execution stage** or **initiation of response execution stage** (Ivry, Franz, Kingstone & Johnston, 1998; Logan & Burkell, 1986; Meyer & Kieras, 1997; but see Nickerson, 1965; Schubert, 1999), demonstrating that two motor production processes cannot be conducted concurrently, even though the response modality of the two tasks is different (e.g. one task requires a verbal response while the other requires a manual press response). A third possibility is that the bottleneck exists at the **response selection stage** (Pashler & Johnston, 1989; Welford, 1952) and fulfilment of the response selection can only be conducted one at a time. Welford (1952) first proposed that decision processes and the response selection stage constitute a single-

channel bottleneck. This means that the response selection of T2 cannot operate at the same time as the response selection of T1, and has to wait until the selection of T1 is completed. Thus, the response time to the second stimulus is postponed with increasing overlap of T1 and T2. Pashler and Johnston (1989) tested the assumptions of whether the bottleneck is located at the response selection stage by manipulating the duration of each processing stage, since, fortunately, different assumptions which suggest the different bottleneck location predict very distinct results, as discussed in the following sections.

#### 2.4.2.1.1 Manipulating the duration of the perceptual stage

The duration of the perceptual stage can be manipulated by using stimulus contrast/intensity (Pashler, 1984) or display size (De Jong, 1993; Pashler, 1984). Stimulus quality impacts on the visual encoding, such that high stimulus quality/intensity (e.g. presenting a white word *Hello* on a black background) shortens the perceptual stage compared to low stimulus quality/intensity (e.g. presenting a grey word *Hello* on a black background). An alternative way of manipulating the perceptual stage is display size (Johnsen & Briggs, 1973). For example, an increase in the stimuli display size corresponds to an increase in the time needed for visual searching (see Pashler, 1984), and therefore, prolong the duration before the response selection stage. So, in a task requiring subjects to decide whether a particular letter (e.g. the letter G) is in a letter array, displaying a six-letter array (e.g. letter array R O D Y G A) is more time-consuming for completing the perceptual stage compared to displaying a two-letter array (e.g. letter array N X).

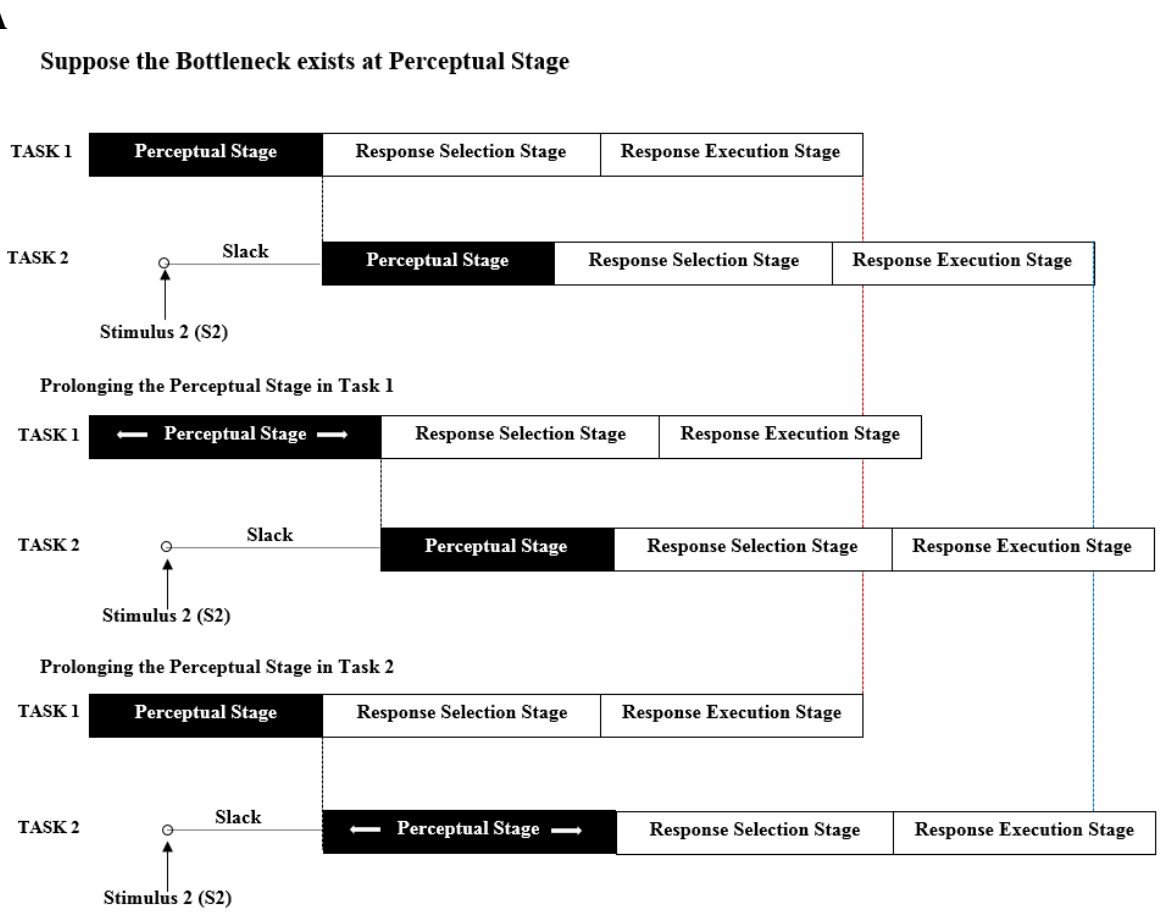
***If the bottleneck is located at the perceptual stage***, then both the reaction time to the first task (RT1) and to the second task (RT2) are postponed when the perceptual stage of T1 is prolonged. Prolonging the perceptual stage in T1 increases the RT1 accordingly. At the same time, prolonging the perceptual stage in T1 increases the slack between the perceptual stage and response selection stage in T2, because the perceptual stage in T2 cannot begin until the perceptual stage in T1 is finished due to the perceptual bottleneck. Thus, increasing the perceptual stage of T1 also increases the RT2. If the perceptual stage in T2 is prolonged, then RT1 is uninfluenced while RT2 is prolonged (see Fig. 2.15).

***If the bottleneck is located at the response selection stage,*** then both RT1 and RT2 increase when the perceptual stage of T1 is prolonged (see Fig. 2.15). Increasing the

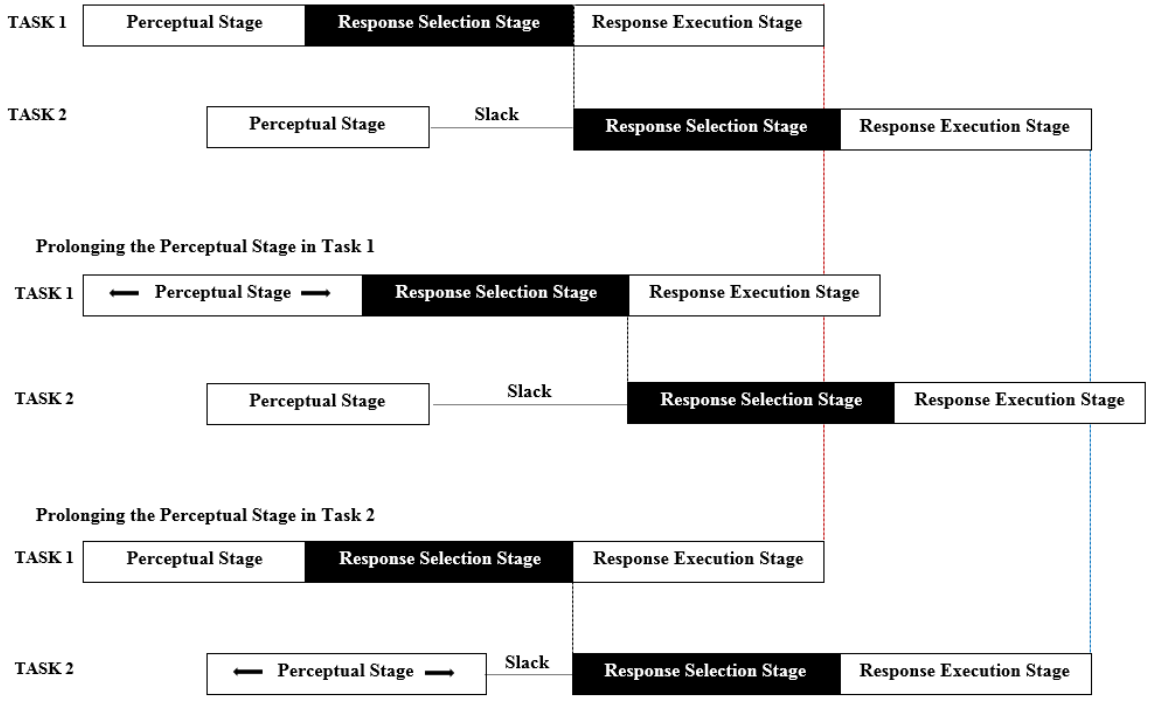
duration of the perceptual stage in T1 postpones its subsequent stages (the response selection stage and response execution stage) and accordingly, prolongs its reaction time. Since the response selection stage in T1 is postponed, the same stage in T2 has to be postponed as well because it has to wait until the response selection stage in T1 is finished, while the perceptual stage in T2 is relatively un-influenced. Thus, the slack between the perceptual stage and response selection stage in T2 increases and prolongs the RT2. However, when the duration of the perceptual stage in T2 is increased, the response selection model predicts that both RT1 and RT2 will remain unchanged. Since the bottleneck is located at the response selection stage, the response selection stage in T2 cannot begin before the completion of the same stage in T1, instead of waiting for the completion of its previous stage in T2 at short SOAs. Therefore, increasing the duration of the perceptual stage in T2 cannot prolong the RT2 (compared to the condition where the perceptual stage is not prolonged) because the prolonged perceptual stage will be "absorbed" into the slack stage.

***If the bottleneck is located at the response execution stage,*** then the consequences of prolonging the perceptual stage in T1 or T2 are the same as the prediction of the response selection bottleneck model (see Fig. 2.15). Both RT1 and RT2 increase when prolonging the perceptual stage in T1, while both of them remain the same when increasing the duration of the perceptual stage in T2. The effect of prolonging the pre-bottleneck stage(s) only occurs at long SOA and is "absorbed" at short SOA as suggested above.

**A**


Suppose the Bottleneck exists at Perceptual Stage
TASK 1
Perceptual Stage
Response Selection Stage
Response Execution Stage
TASK 2
Slack
Perceptual Stage
Response Selection Stage
Response Execution Stage
Stimulus 2 (S2)
Prolonging the Perceptual Stage in Task 1
TASK 1
Perceptual Stage
Response Selection Stage
Response Execution Stage
TASK 2
Slack
Perceptual Stage
Response Selection Stage
Response Execution Stage
Stimulus 2 (S2)
Prolonging the Perceptual Stage in Task 2
TASK 1
Perceptual Stage
Response Selection Stage
Response Execution Stage
TASK 2
Slack
Perceptual Stage
Response Selection Stage
Response Execution Stage
Stimulus 2 (S2)

**B**


Suppose the bottleneck exists at Response Selection Stage
TASK 1
Perceptual Stage
Response Selection Stage
Response Execution Stage
TASK 2
Perceptual Stage
Slack
Response Selection Stage
Response Execution Stage
Prolonging the Perceptual Stage in Task 1
TASK 1
Perceptual Stage
Response Selection Stage
Response Execution Stage
TASK 2
Perceptual Stage
Slack
Response Selection Stage
Response Execution Stage
Prolonging the Perceptual Stage in Task 2
TASK 1
Perceptual Stage
Response Selection Stage
Response Execution Stage
TASK 2
Perceptual Stage
Slack
Response Selection Stage
Response Execution Stage

**C**

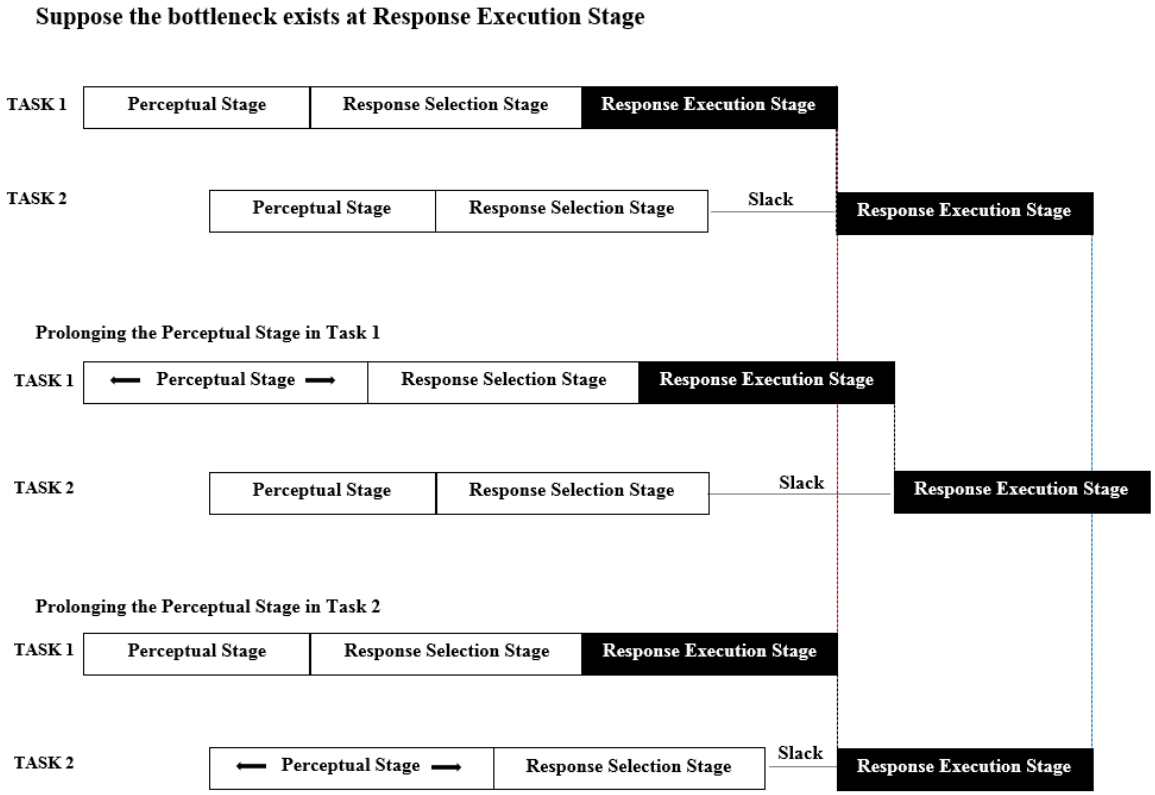



**Figure 2. 15** Assumption that the bottleneck exists at perceptual stage (A), response selection stage (B), or response execution stage (C). The black box represents where the bottleneck is located, while the white box represents the stage that can be conducted concurrently with other stages. The small circle in A represents the time that the second stimulus is shown. The first graph of A, B and C indicates the baseline of the dual task. The second graph shows the situation of increasing the duration of the perceptual stage in Task 1, while the third presents the structure of prolonging the perceptual stage in Task 2.

#### 2.4.2.1.2 Manipulating the duration of the response selection stage

The duration of the response selection stage can be manipulated by stimulus repetition (Bertelson, 1963; Pashler & Johnston, 1989) or by changing stimulus-response (S-R) binding difficulty (Pashler, 1989). Repeating the same T2 stimulus from the previous trial eases the mental retrieval of S-R binding (e.g. in trial 1, T2 presents a letter "A" which requires the participant to press a red button; in trial 2, the same stimulus letter "A" is presented for which the corresponding response is also pressing a red button – however, T1 in trial 1 and trial 2 are independent). Increasing the difficulty of S-R mappings can also increase the duration of the response selection stage accordingly. For example, difficult S-R mappings (e.g. requiring subjects to press button 2, 3, and 1 in response to low, medium

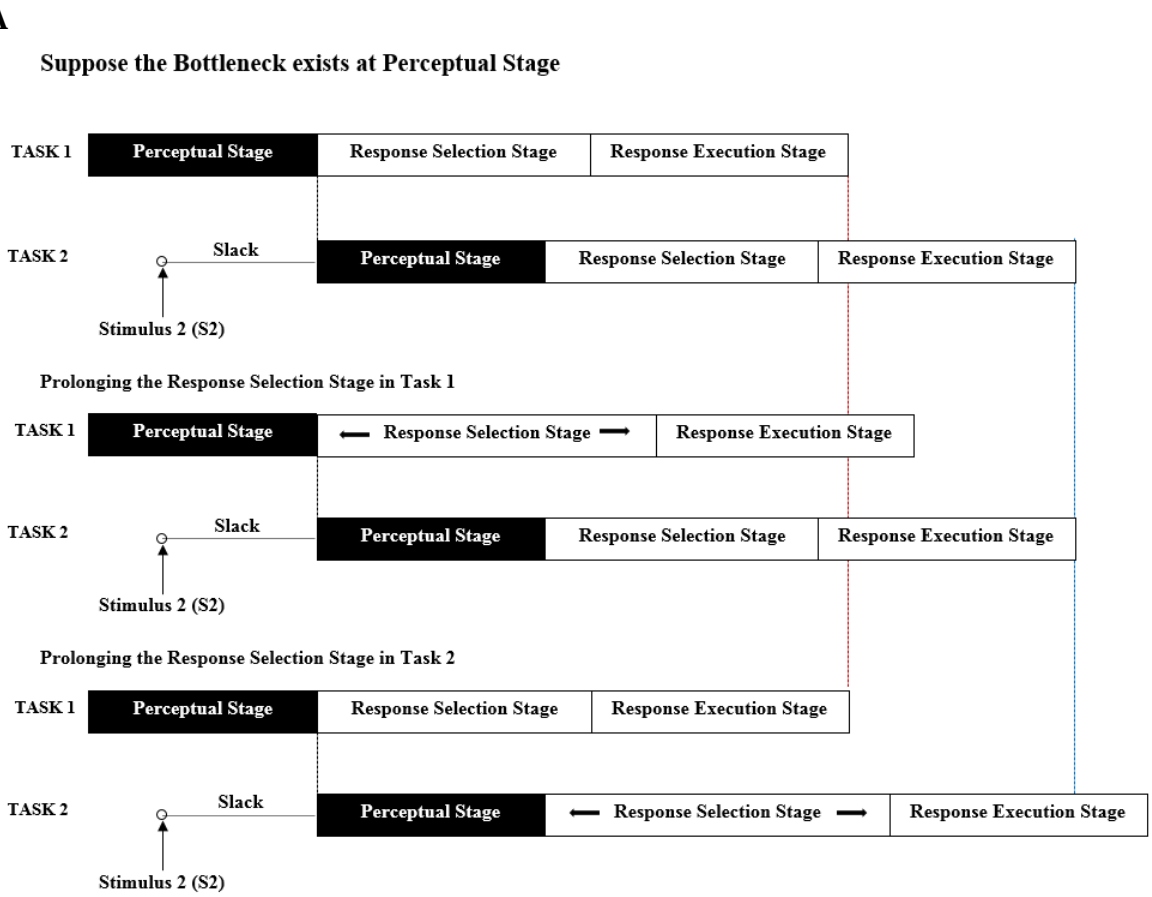
and high pitch tones, respectively) require more time in the response selection stage compared to easy S-R mappings (e.g. requiring subjects to press 1, 2, and 3 in response to low, medium and high pitch tones, respectively).

***If the bottleneck is located at the perceptual stage***, then increasing the response selection stage in either T1 or T2 can only increase its corresponding RT but has no interference on RT in another task since the response selection stage is post-bottleneck stage, and can be conducted concurrently with other stage(s) (see Fig. 2.16).

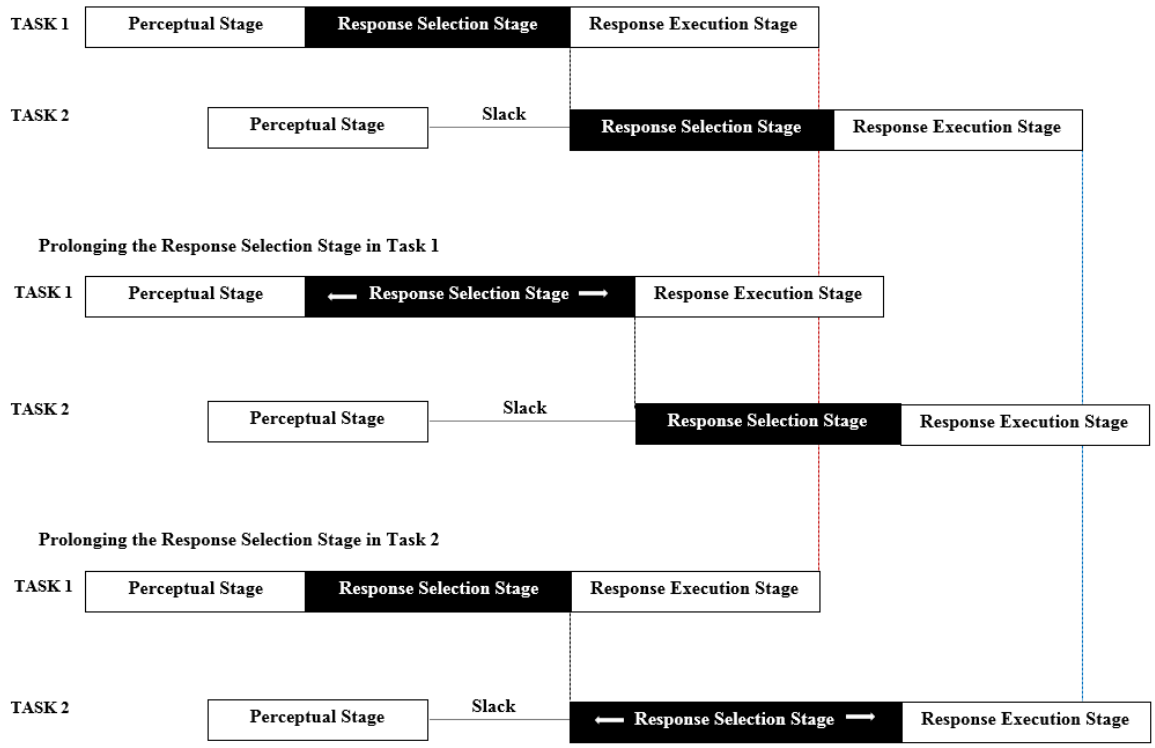
***If the bottleneck is located at the response selection stage,*** and the response selection stage in T1 is increased, the response selection (bottleneck) stage in T2 cannot commence until the prolonged bottleneck stage in T1 is complete. The slack before the bottleneck stage in T2 increases accordingly. As a consequence, both RT1 and RT2 are prolonged to the same extent. If the bottleneck stage in T2 is increased, then RT1 should not be impacted because it has already finished before the response selection stage in T2 begins. Thus, only RT2 is prolonged (see Fig. 2.16).

***If the bottleneck is located at the response execution stage,*** increasing the stage(s) previous to the bottleneck stage postpones the bottleneck and its later stage(s). Therefore, increasing the response selection stage in T1 postpones its later stage and prolongs RT1. RT2 is prolonged as well due to the increment of the slack. Both RT1 and RT2 increase with prolonging the response selection stage in T1. However, increasing the duration of the response selection stage in T2 has no impact on either RT1 or RT2, because of the underadditive interaction with the previous bottleneck stage increment (see Fig. 2.16).

**A**


Suppose the Bottleneck exists at Perceptual Stage
TASK 1
Perceptual Stage
Response Selection Stage
Response Execution Stage
TASK 2
Slack
Perceptual Stage
Response Selection Stage
Response Execution Stage
Stimulus 2 (S2)
Prolonging the Response Selection Stage in Task 1
TASK 1
Perceptual Stage
Response Selection Stage
Response Execution Stage
TASK 2
Slack
Perceptual Stage
Response Selection Stage
Response Execution Stage
Stimulus 2 (S2)
Prolonging the Response Selection Stage in Task 2
TASK 1
Perceptual Stage
Response Selection Stage
Response Execution Stage
TASK 2
Slack
Perceptual Stage
Response Selection Stage
Response Execution Stage
Stimulus 2 (S2)

**B**


Suppose the bottleneck exists at Response Selection Stage
TASK 1
Perceptual Stage
Response Selection Stage
Response Execution Stage
TASK 2
Perceptual Stage
Slack
Response Selection Stage
Response Execution Stage
Prolonging the Response Selection Stage in Task 1
TASK 1
Perceptual Stage
Response Selection Stage
Response Execution Stage
TASK 2
Perceptual Stage
Slack
Response Selection Stage
Response Execution Stage
Prolonging the Response Selection Stage in Task 2
TASK 1
Perceptual Stage
Response Selection Stage
Response Execution Stage
TASK 2
Perceptual Stage
Slack
Response Selection Stage
Response Execution Stage

**C**

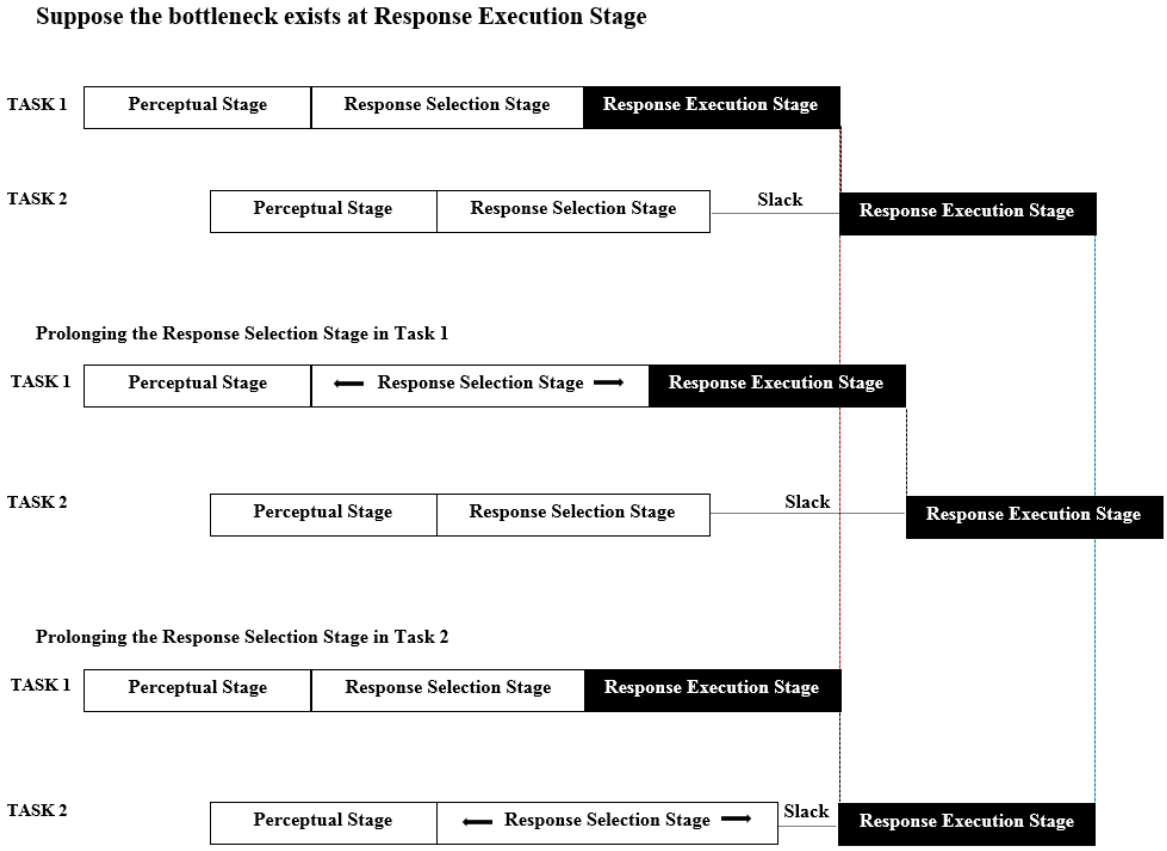



**Figure 2. 16** Assumption that the bottleneck exists at perceptual (A), response selection (B), or response execution stage (C). The black box indicates the bottleneck location, while the white box shows the stage that can overlap with other stages. The small circle in A indicates the presentation of the second stimulus. The first graph of A, B and C indicates the baseline, followed by the situation of increasing the response selection stage duration in Task 1 and in Task 2.

### 2.4.2.1.3 Manipulating the duration of the response execution stage

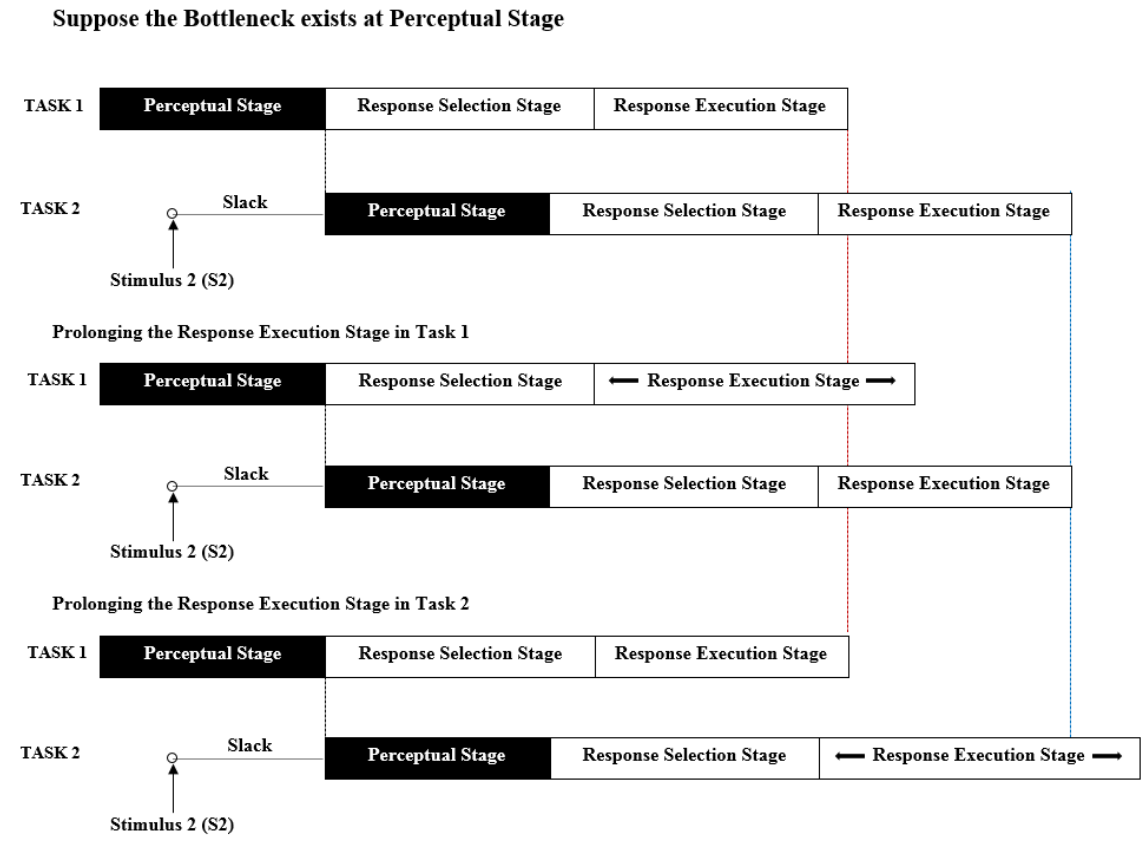
Increasing the complexity of producing a selected response is a way of manipulating the duration of the response execution stage (Pashler & Christian, 1994). For example, requiring subjects to press three sequential buttons (e.g. the three adjacent keys J, K, L on the keyboard) instead of just one (e.g. the key J) in response to a stimulus requires more time to complete the response. However, the reaction time to press the first of the three buttons is the same as just pressing one, meaning that this manipulation impacts on the response execution stage but not the response selection stage.

***If the bottleneck is located at the perceptual stage***, then increasing the post-stage (response execution stage) does not have an impact on the performance of the other task since the non-bottleneck stage(s) can be conducted at the same time with another task. Therefore, increasing the response execution stage in T1 can only prolong the RT1, while manipulating the duration of the response execution stage in T2 only has an impact on the performance of RT2 (see Fig. 2.17).

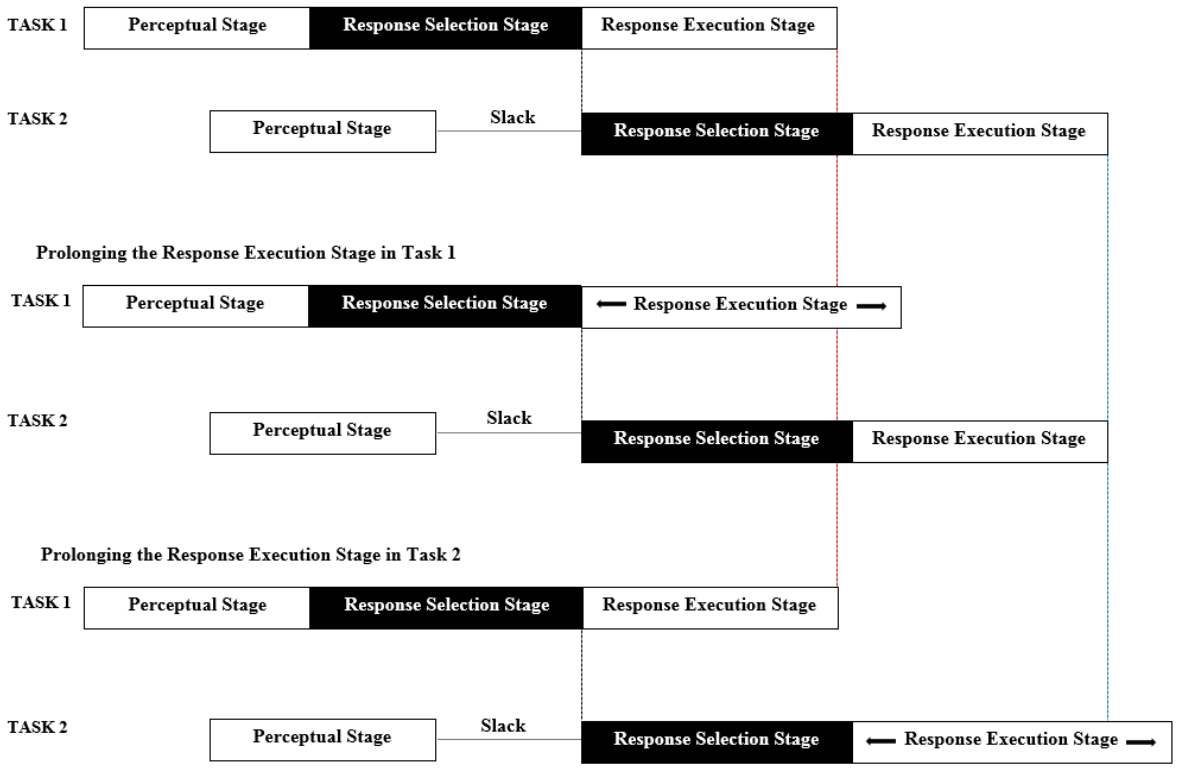
***If the bottleneck is located at the response selection stage,*** then manipulating the duration of the response execution stage implies the same prediction as for the perceptual bottleneck assumption, above, because the response execution stage occurs after the bottleneck stage. Manipulating the post-bottleneck stage only impacts on its own reaction time, while the performance of the other task is fairly independent of the manipulation (see Fig. 2.17).

***If the bottleneck is located at the response execution stage,*** increasing the bottleneck stage in T1 not only impacts on T1 performance but also impacts on the performance of T2, because the bottleneck stage of T2 has to wait until the completion of the increased bottleneck stage in T1. Thus, RT1 and RT2 increase to the same extent when prolonging the duration of the response execution stage in T1. However, only RT2 is prolonged when increasing the duration of the response execution stage in T2 because the performance of T1 is finished before the bottleneck stage in Task 2 starts (see Fig. 2.17).

**A**


Suppose the Bottleneck exists at Perceptual Stage
TASK 1
Perceptual Stage
Response Selection Stage
Response Execution Stage
TASK 2
Slack
Perceptual Stage
Response Selection Stage
Response Execution Stage
Stimulus 2 (S2)
Prolonging the Response Execution Stage in Task 1
TASK 1
Perceptual Stage
Response Selection Stage
Response Execution Stage
TASK 2
Slack
Perceptual Stage
Response Selection Stage
Response Execution Stage
Stimulus 2 (S2)
Prolonging the Response Execution Stage in Task 2
TASK 1
Perceptual Stage
Response Selection Stage
Response Execution Stage
TASK 2
Slack
Perceptual Stage
Response Selection Stage
Response Execution Stage
Stimulus 2 (S2)

**B**


Suppose the bottleneck exists at Response Selection Stage
TASK 1
Perceptual Stage
Response Selection Stage
Response Execution Stage
TASK 2
Perceptual Stage
Slack
Response Selection Stage
Response Execution Stage
Prolonging the Response Execution Stage in Task 1
TASK 1
Perceptual Stage
Response Selection Stage
Response Execution Stage
TASK 2
Perceptual Stage
Slack
Response Selection Stage
Response Execution Stage
Prolonging the Response Execution Stage in Task 2
TASK 1
Perceptual Stage
Response Selection Stage
Response Execution Stage
TASK 2
Perceptual Stage
Slack
Response Selection Stage
Response Execution Stage

**C**

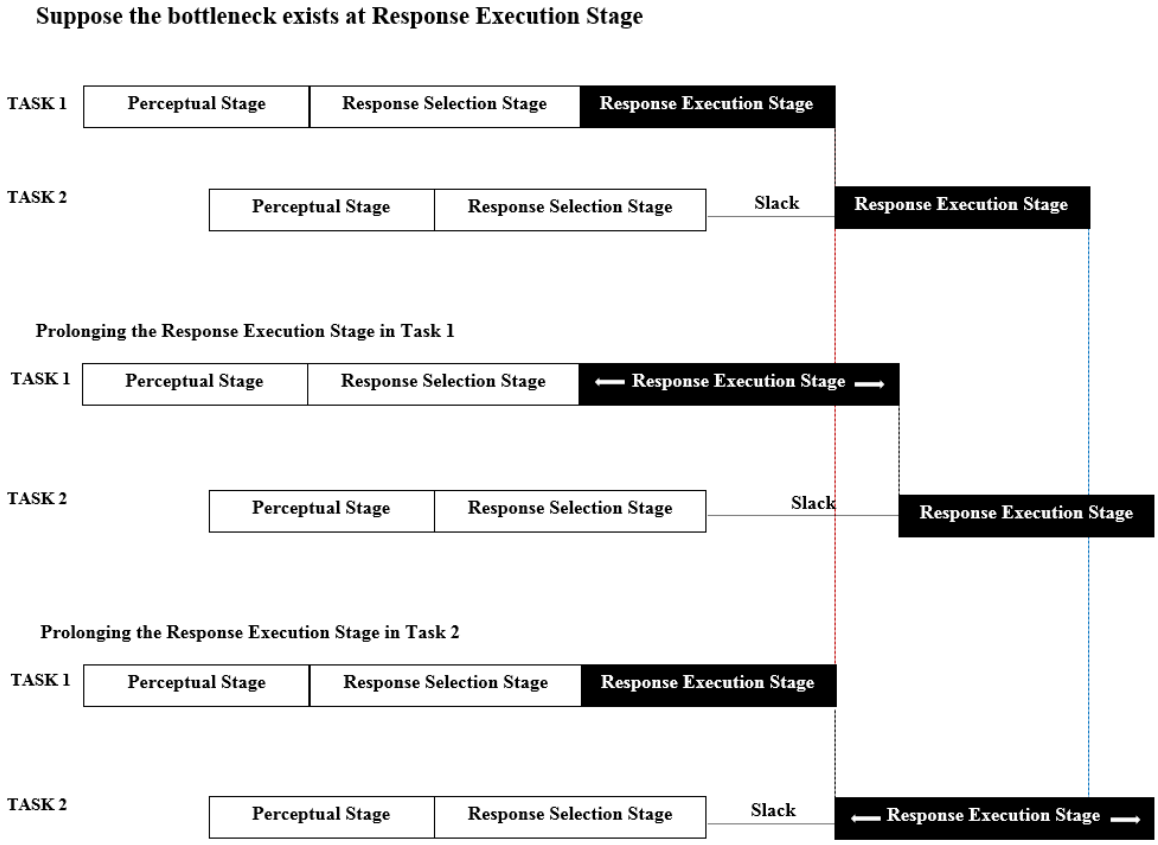



**Figure 2. 17** The assumption that the bottleneck exists at perceptual (A), response selection (B), or response execution stage (C). The black box indicates the bottleneck location, while the white box shows the stage that can overlap with other stages. The small circle in A indicates the presentation of the second stimulus. The first graph of A, B and C indicates the baseline, followed by the situation of increasing the response selection stage duration in Task 1 and in Task 2.

The serial studies of Pashler and Johnston (1989) manipulating the duration of the three stages yield the following findings:

1. When increasing the duration of the perceptual stage in T2, RT2 is un-impacted and not postponed. This result provides evidence against the perceptual bottleneck assumption, which predicts that RT2 will be postponed. This result is consistent with both the response selection bottleneck and the response execution bottleneck assumption, which entail that pre-bottleneck manipulation in T2 will be absorbed into the slack and performance will remain the same.
2. When prolonging the response selection stage of T1, both RT1 and RT2 increase to the same extent. This result provides evidence against the prediction of an un-influenced RT2 of the perceptual bottleneck model but favours the prediction of both the response

selection bottleneck and the response execution bottleneck model. However, when increasing the duration of the response selection stage in T2, RT2 is prolonged accordingly. This result provides evidence against the inference of the response execution bottleneck model, which suggests that RT2 remains the same when manipulating the pre-bottleneck stage, but is consistent with the response selection bottleneck model.

3. When increasing the duration of the response execution stage in T1, RT2 is independent of this manipulation and remains unaffected. The result is consistent with the prediction of the response selection bottleneck model but is evidence against the response execution bottleneck, which suggests that RT2 is influenced by manipulating the bottleneck stage in T1.

In sum, the research results provide strong support for the response selection bottleneck model (Pashler & Johnston, 1989). Pashler (1989) proposed the **two-component model** (see Fig. 2.18) which suggests there are two points in processing where interference takes place: at the **perceptual processing stage** and the **response selection stage**. As shown in Figure 2.18, perceptual processing commences immediately after each stimulus arrives. If the stimuli are difficult to process in the perceptual processing stage and exceed a certain level of difficulty (shown as the width of the channel which contains two semi-circles), then the first interference occurs and **degrades the accuracy** of both processing stimuli. After finishing encoding the first stimulus, the central mechanism will be occupied and begins to select the appropriate response. If the second stimulus is processed before the response selection on T2 is complete (which occurs in the short SOA), **postponement** occurs due to queueing at the response selection stage.

The indication of interference which occurs in the perceptual stage comes from the results of multi-stimuli task experiments. The accuracy of a multi-stimuli task is impaired when perceptual processing difficulty (manipulated by increasing the discrimination difficulty or enlarging the stimuli display size) surpasses a certain degree. An example of this is when a visual discrimination task is made more complex by presenting lots of red and green letter Rs and red letter Ts, and requiring participants to search for the unique green letter T (see Kleiss & Lane, 1986; Neisser, Novick & Lazar, 1963; Treisman & Gelade, 1980). The interference which occurs at the response selection stage, as shown above, comes from the multi-tasks and is caused by an inability to conduct two response selection stages at the

same time. However, the question arises whether it is possible that the dual task interference might be caused by other reasons, such as exceeding some capacity limitation (similar to the reason for interference at the perceptual stage), instead of being caused by a structural limitation as the response selection bottleneck suggests. The following section illustrates another model which provides a different assumption accounting for the PRP effect.

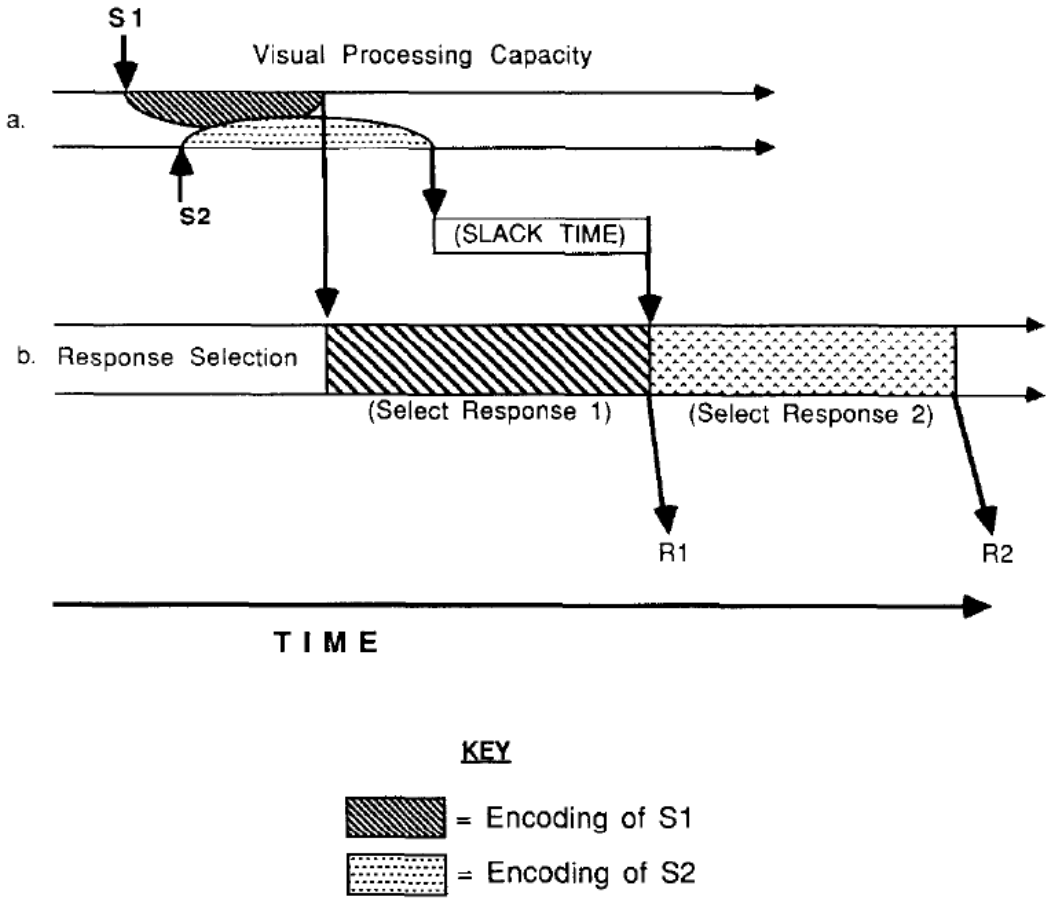



**Figure 2. 18** The two-component model, illustrating interference at two points in processing: simultaneous degradation of perceptual processing, and queueing at response selection. From Pashler, H. (1989). Dissociations and dependencies between speed and accuracy: Evidence for a two-component theory of divided attention in simple tasks. *Cognitive Psychology*, *21*(4), 469-514.

### 2.4.2.2 The capacity sharing model

In contrast to the bottleneck model, the capacity sharing model (e.g. Kahneman, 1973; Mcleod, 1997) favours the assumption that the processing stages can be conducted in parallel. The model suggests that there are limited cognitive/mental resources or capacity,

and these resources are shared by all the tasks that are conducted concurrently. When the total mental capacity is sufficient for the *workload* (the amount of mental capacity required to solve the task(s)) of the two/multiple tasks, then the tasks can be conducted simultaneously without any interference. However, when the workload of the two/multiple tasks exceeds the total amount of mental resources, then interference occurs and reaction time increases accordingly, due to the decreased amount of capacity allocation to each task (Allport, Antonis, & Reynolds, 1972).

In dual tasks, the limited capacity is shared by two successive tasks. When decreasing the SOA, the overlapping processing stages in two tasks increase. The duration of the processing stage(s) allocated less resources in each task increases accordingly; reaction time is therefore prolonged and accuracy is impacted (see Fig. 2.19). This also implies that manipulating the capacity requirement in overlapping stage(s) while keeping the SOA constant can also lead to an increased workload requirement, and when the requirement surpasses the total limited capacity, the performance is impacted as a consequence.

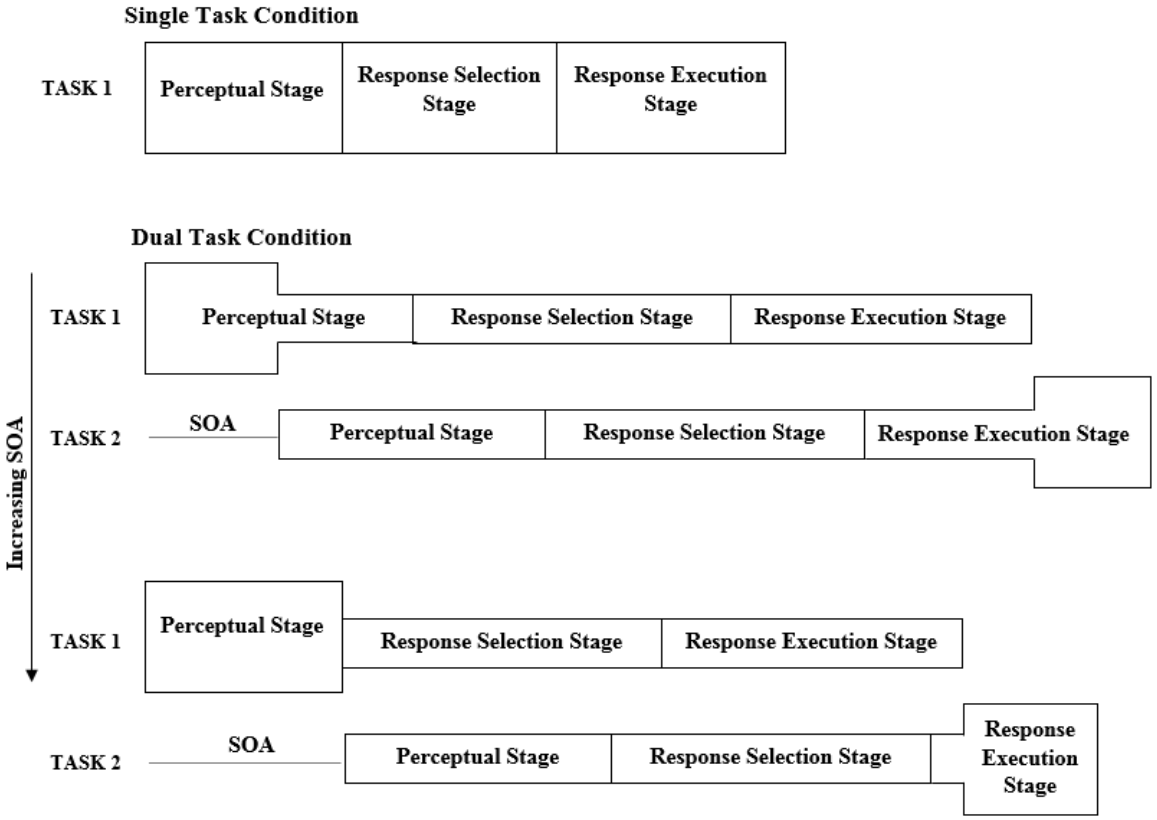



**Figure 2. 19** The typical capacity sharing model. The height of the box represents the limited capacity that is allocated to the task. The width of the boxes represents the reaction time to the task. In the single task condition, capacity is (almost) fully allocated to the task. In the dual task condition, the capacity is first (almost) fully allocated to T1 before T2 begins. After T2 starts, the capacity is shared by two tasks and each task receives limited capacity. The longer the overlapping stages are, the longer the reaction time will be.

One of the objections raised to this capacity sharing model is based on findings showing that when increasing the duration of the perceptual stage in T2, RT2 remains the same, instead of increasing accordingly (Pashler & Johnston, 1989). Moreover, when a less effortful visual detection task (e.g. detecting the specific (visual) signal by raising the hand) is combined with a more effortful auditory detection task, its performance is less impacted, compared to combining it with another less effortful visual detection task. This suggests that increasing the amount of capacity required by combining an easier task with a more difficult task (rather than a task of similar difficulty) under dual task conditions may not lead to worse performance. These results are taken as evidence against the capacity sharing model (Segal & Fusella, 1970). One possible explanation that might account for this result while still favouring the capacity sharing assumption could be that each cerebral hemisphere (right and left) has its limited capacity (Dimond, 1970, Friedman, Polson, &

Dafoe, 1988). If two tasks require the limited capacity from the same hemisphere, such as requiring the same input and output modality (e.g. two visual detection tasks), interference occurs and performance is impaired because of the inadequate capacity that each task receives. When two tasks require different resources from each hemisphere (e.g. one task requires visual input while another one requires auditory input) dual task interference can be reduced or even eliminated. However, research results show that interference can still be obtained when each task is carried out by each hemisphere (e.g. a verbal response is controlled by the left hemisphere, while a left-hand manual response is controlled by the right hemisphere), arguing against the assumption that dual task interference can be avoided when two tasks require the limited capacity from different hemispheres (Pashler & O'Brien, 1993).

Supporting evidence for capacity sharing models is found in findings that RT2 is prolonged in short SOA conditions (e.g. 50 ms), while RT1 is delayed and produced at the same time as or close to RT2 at long SOA conditions (e.g. 900 ms) (Kahneman, 1973; also see Pashler, 1984), suggesting that the limited capacity is shared by two tasks, and either RT1 or RT2 is impacted under dual task conditions due to the depleted capacity they have been allocated (see Fig. 2.20). These results are used to argue against the bottleneck model since, according to the bottleneck model, T1 is normally completed before the presentation of T2 at long SOA conditions in most of the typical dual task paradigm studies. In other words, the bottleneck stages of the two tasks do not overlap and the performance of the two tasks should be as short as when it is conducted in the single task condition.

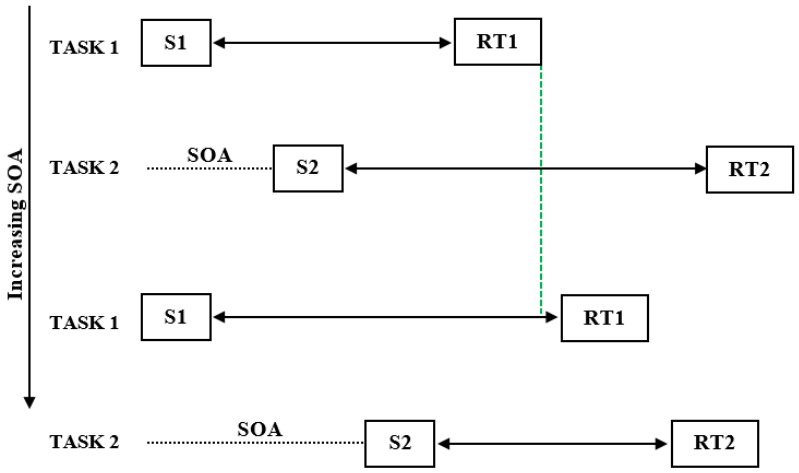



**Figure 2. 20** The phenomenon of decreasing RT2 at short SOA while prolonging RT1 at long SOA in the dual task. S1 and RT1 represent the stimulus and reaction time to Task 1, respectively; while S2 and RT2 represent the stimulus and reaction time to Task 2, respectively. Two tasks are presented in succession. When the interval between the two tasks is short (as in the graph on top), RT2 is normally prolonged, compared to the reaction time at long SOA. In contrast, in the long SOA condition (as in the graph at the bottom), RT1 is normally increased.

However, adherents of the bottleneck model point out that this kind of result does not contradict the bottleneck model, and can be explained logically (Pashler & Johnston, 1989). They suggest that some participants adopt a strategy of holding back the production of T1 after selecting the response, and produce the two responses at the same time or in rapid succession after finishing the response selection of T2 (See Fig. 2.21). Therefore, the slowing of RT1 when the SOA is long is the consequence of this strategy, and not of dual-task interference caused by capacity sharing. This type of strategy has been termed a "grouping" strategy (Borger, 1963). This explanation has been tested by encouraging participants to "group" their responses, or conjoin the responses. The inter-response intervals (IRIs) (the interval between RT1 and RT2) should remain substantial after combining or grouping the responses, regardless of SOA manipulation. Moreover, the stage manipulation of T2 should pass to both RT1 and RT2 to the same extent if the "grouping" strategy is adopted. Research results have shown that the manipulation of the perceptual stage in T2 has no impact on either RT1 or RT2, while an increase in the duration of the response selection stage in T2 leads to the result of prolonging both RT1 and RT2 to the same extent (Pashler & Johnston, 1989). These results support the inference of a "grouping" strategy and explain the RT1 slowing in line with the bottleneck model.

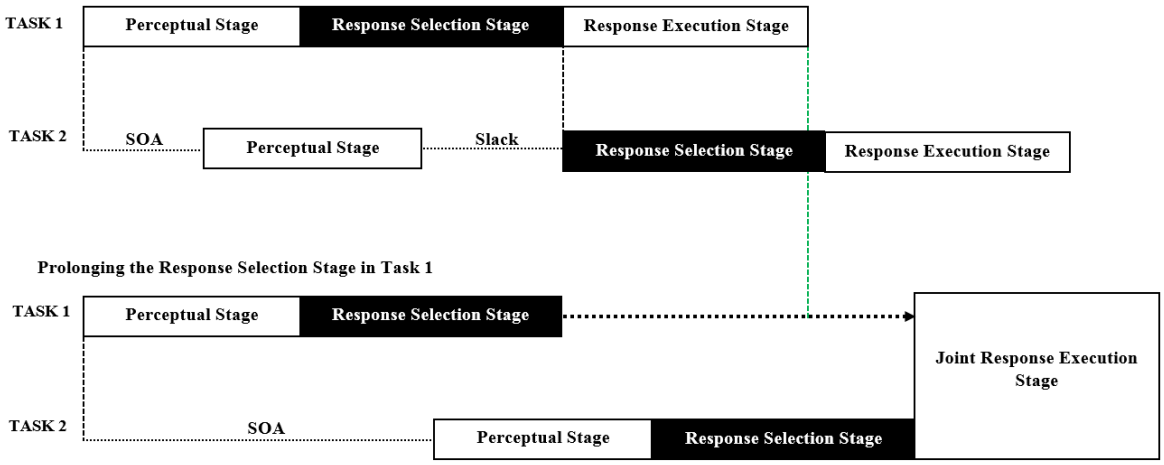



**Figure 2. 21** The top graph represents the sequence of dual task processing at short SOA, while the bottom presents the processing sequence of the response "grouping" strategy at long SOA. The response selection stage of the two tasks is conducted in serial, and the response of Task 1 is held back and executed with the response of Task 2.

Another proposal that has been made by some researchers is that the dual task interference (or PRP) effect is an artefact that is a consequence of the emphasis on the priority of RT1 in the instruction (Meyer et al., 1995; Ruthruff, Miller & Lachmann, 1995). It has been proposed that cognitive capacity can be allocated in a differentiated way, in accordance to the features of tasks, and can also be allocated voluntarily (McLeod, 1977). In this view, in dual task experiments participants allocate (almost) all of their capacity on T1 first, and then on T2, because of the meaning they take from the instruction. As a consequence, RT1 is un-impacted and RT2 is prolonged when the SOA is decreased. Thus, dual task interference is the consequence of strategic postponement rather than a structural limitation, as is proposed by the bottleneck model. Compelling support for this argument comes from the study of Schumacher and colleagues (2001), showing that interference in a dual task can be eliminated after almost 2000 trials by emphasising the importance of the two tasks equally. The speed and accuracy of RT1 and RT2 are similar when conducted in the dual task condition and when conducted individually.

However, Levy and Pashler (2001) point out that the reason why most researchers emphasise the priority of RT1 is to avoid inducing a "grouping" strategy. Moreover, they conducted several studies mimicking that of Schumacher and colleagues (2001), emphasising the importance of the two tasks equally. Their results show that a substantial

PRP effect can still be obtained. Levy and Pashler (2001) suggest that the results showing parallel responses in the study of Schumacher and colleagues (2001) are the consequence of the highly compatible S-R binding used in the study (e.g. press the button with the index, middle and ring figure to the stimulus O when it presents on left, middle, and right of the screen respectively; i.e., press the most-left key corresponding to most-left position and so on). The more compatible the S-R mapping is, the shorter the response selection stage is. Furthermore, dual task interference can still be obtained even when equally emphasising the importance of two tasks and not stipulating the response order (Ruthruff, Pashler & Hazeltine, 2003).

Other evidence against the capacity sharing model is provided by Ruthruff, Pashler and Klaassen (2001). They tested the bottleneck model and capacity sharing model by manipulating the duration of the response selection stage of T1. They instructed their participants to divide their attention between the two tasks equally and encouraged them to group their responses to the simultaneous tasks. When the response selection of T1 is easy (see Fig. 2.22), the response execution has to be held back even if its processing has been completed. When the difficulty of the first task is increased to some extent, the reaction time of the grouped responses should remain the same according to the capacity sharing model, since the grouped reaction time depends on the slower response (e.g. T2), and RT2 should not be prolonged because T2 is still allocated (almost) half of the capacity regardless of the manipulation of T1. The results show that the grouped responses were prolonged when increasing the difficulty of T1, which is consistent with the prediction of the bottleneck model (see Fig. 2.22).

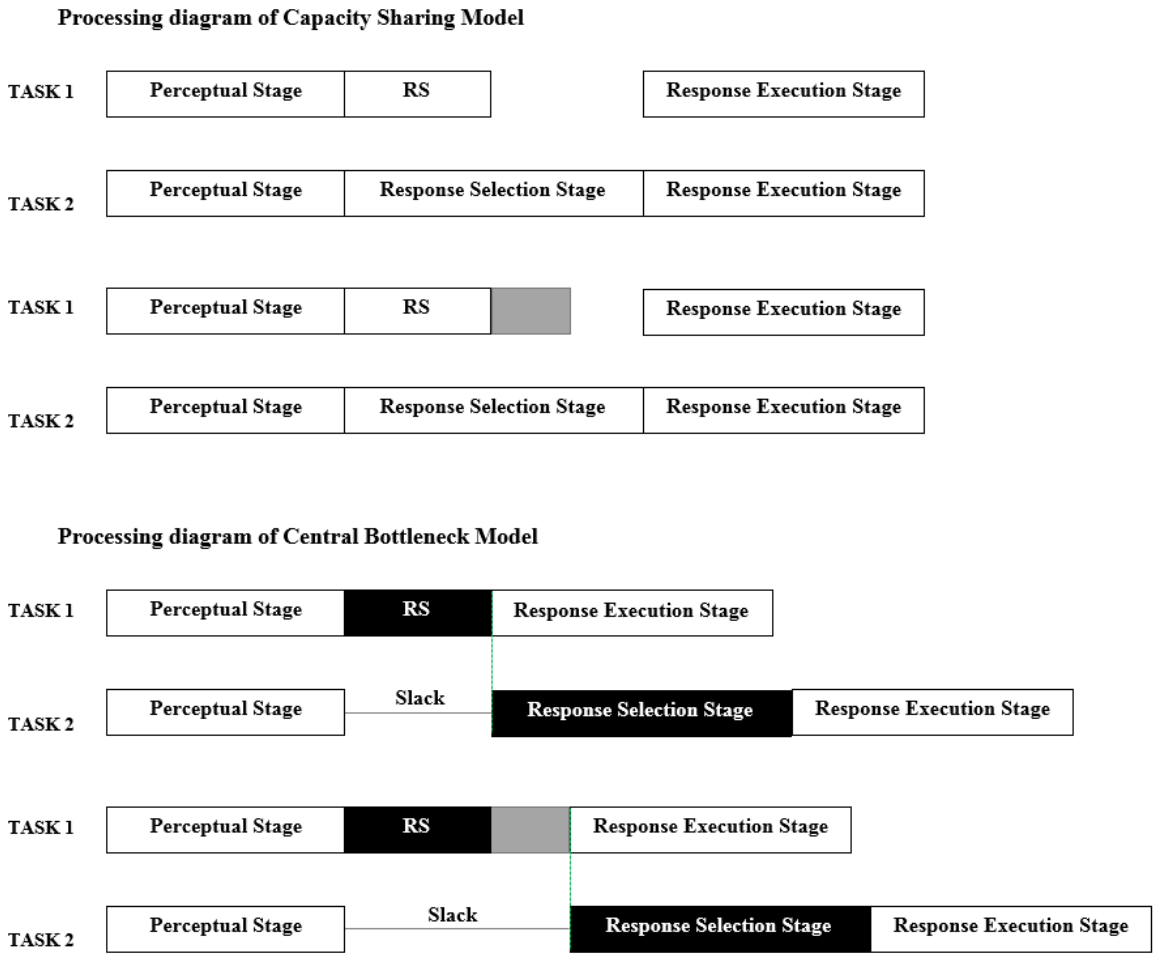



**Figure 2. 22** The predictions of the capacity sharing model and the central bottleneck model when manipulating the response selection stage. The top part within each model presents the baseline while the bottom part presents the processing sequence after manipulating the duration of the response selection stage. The grey box indicates the increased duration of the response selection. The black box represents the central processing bottleneck. RS presents the response selection stage.

An alternative model has been proposed to account for the empirical results that seem to contradict the assumption that limited capacity is shared by all the processing stages. This model suggests that capacity sharing is restricted to a central stage, the response-selection stage (Tombu & Jolicœur, 2003). This model may be termed the **central capacity sharing model**. This model shares the essential assumption of the capacity sharing model, namely that the processing of two/multiple tasks can be conducted concurrently by sharing limited capacity. The processing speed of the task will be slowed down when sharing parts of the capacity under a two/multiple task condition, compared to when the full capacity allocation is received under the single task condition. However, in contrast to the capacity sharing model outlined above, this model assumes that instead of all the processing stages, only the

central stage (response-selection stage) is subject to capacity sharing. Contrary to the response selection bottleneck model, the response selection stage of two tasks can be conducted in parallel but not in serial.

According to the central capacity sharing model (which supposes that the limited capacity is shared equally by the central processing stage of two tasks), when increasing the perceptual stage of T2 under a short SOA condition, the overlapping central stage of the two tasks decreases, and therefore, the processing stage allocated less processing capacity decreases as well (see Fig. 2.23). As a result, RT1 decreases because its central stage, which shares the limited capacity with another task, is shorter than before. For RT2, the perceptual stage increases, and its central capacity sharing stage decreases accordingly. These two effects should counteract each other, and thus, the reaction time of T2 remains the same. This inference is consistent with the result of underadditive interaction when manipulating the perceptual stage of T2, and offers an account for one of the strongest arguments against the capacity sharing model.

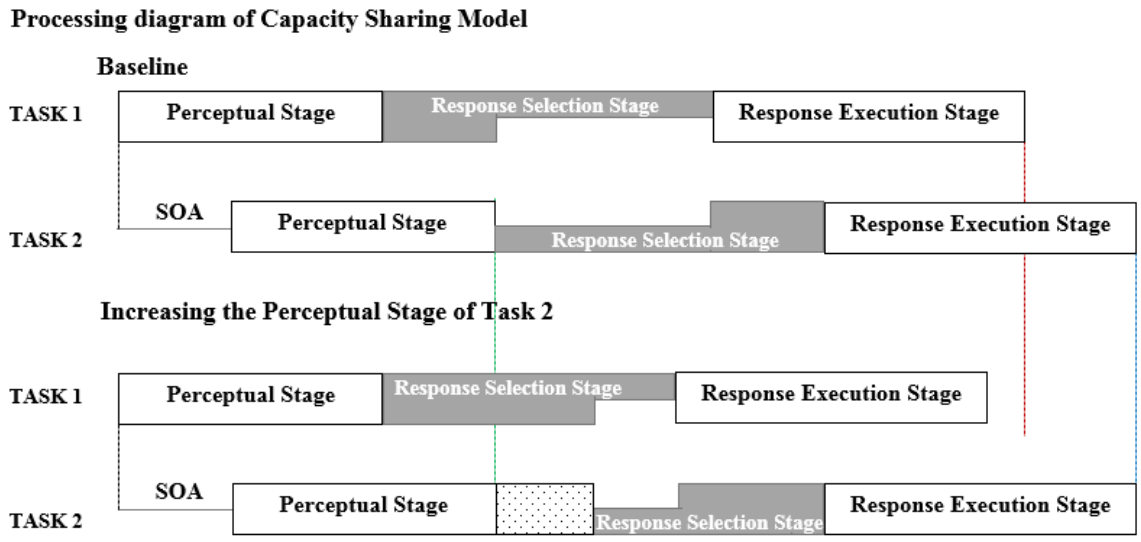



**Figure 2. 23** The prediction of the central capacity sharing model when increasing the duration of the perceptual stage in Task 2. The top graph is the baseline and the bottom is the processing procedure after stage manipulation. The grey box represents the central stage where the capacity is limited and shared by concurrent tasks. The dot box presents the increased duration of the perceptual stage.

The central capacity sharing model predicts that RT1 should decrease with SOA increasing, in contrast to the response selection bottleneck model which suggests that RT1

should be sustained. Tombu and Jolicœur (2003) accounted for the absence of an SOA effect on RT1 as the result of the fact that subjects allocated (almost) full capacity on T1. The PRP effect can be explained as allocating the full or a large proportion of the capacity on T1, and therefore, RT2 is prolonged when decreasing the SOA due to the limited capacity it has received. Tombu and Jolicœur (2003) claim that the response selection bottleneck model is an extreme case of the central capacity sharing model, suggesting that subjects just allocated all or none of their capacity to a task, instead of sharing it between the tasks. In other words, the bottleneck is just one special way in which tasks may be conducted in a more general central capacity sharing model.

To further explore whether the PRP effect is due to the all-or-none capacity sharing strategy, or due to structural limitations as suggested by the bottleneck model, Ruthruff and colleagues (2003) conducted several experiments without requiring the order of two responses (e.g. not specifying which task response should be offered first), but only emphasising the importance of both tasks. The study further encouraged subjects to equally split their attention on both tasks by mixing the presentation order of the two tasks (e.g. the tone discrimination task could be presented before, at same time or after the letter discrimination task). The aim of this mixed-order design is to avoid participants allocating more capacity (if capacity sharing is possible) on T1, as in the condition in which a certain task is presented before another during the whole block. The results show that while some subjects offered grouping responses, most of them showed clear-cut results consistent with the postulation of the bottleneck mode, regardless of which task's response they offered first.

## 2.5 Conclusion

In sum, most of the current research results obtained in the dual task paradigm by manipulating the processing stage (s) and SOA are consistent with the predictions of the response selection bottleneck model. This suggests that dual task interference is due to structural limitations and the critical processing stage (response selection stage) of two tasks cannot be conducted at the same time, but must occur in serial. However, although the response selection bottleneck model has been strongly favoured, the capacity sharing model cannot be fully ruled out. The modified central capacity sharing model offers an account of the results which do not support the original capacity sharing model, suggesting

that the capacity resources are limited at the central stage (response selection stage) and are shared by all the concurrent tasks. In other words, this means that the central processing stage can be conducted in parallel. This model accounts for the phenomenon that RT1 remains unchanged when RT2 increases with SOA decreasing as the all-or-none capacity allocation strategy subjects have adopted. However, a credible explanation is needed to account for the clear-cut "bottleneck" results when the instruction and the design of the experiment itself encouraged subjects to emphasise the two tasks equally.

The discussion in this section clearly shows that interference results when people perform dual/multiple tasks. A question that follows on this is whether practice improves the ability to perform dual/multiple tasks at the same time, with reduced interference. Studies have shown that extensive practice allows people to conduct two tasks in parallel without any interference (Hazeltine, Teague & Ivry, 2002). The locus of the practice effect is suggested to be at the response selection stage, and the duration of response selection is reduced as a result of practice because the binding of stimulus and response is strengthened by repetition (Pashler & Baylis, 1991; Welford, 1976).

Practice can therefore reduce the duration of the stage that is subject to the central processing bottleneck. Moreover, as mentioned before, language production is subject to the same central processing bottleneck. In the context of this study, this raises the question of whether simultaneous interpreters, who have extensive professional experience in conducting multiple tasks at the same time during simultaneous interpreting, might have comparatively shorter central bottleneck stages during language production compared to monolinguals and also bilinguals since they have received extensive practice at performing dual/multiple tasks. To investigate this question, the methodology outlined in the following chapter was conducted to compare the performance of professional simultaneous interpreters, proficient bilinguals and monolinguals.

# Chapter 3. Methodology

## 3.1 Introduction

This chapter first sets out the research questions and hypotheses that arise from the discussion in Chapter 2 (see Section 3.2). Subsequent to this, a detailed discussion of the experimental design is presented (see Section 3.3), with particular attention to questions of sampling, selection criteria, apparatus and stimuli, design and procedure.

## 3.2 Research questions and hypotheses

Against the background of the literature review presented in the previous chapter, and the gaps identified, this study aims to explore the following research questions:

1) Are professional simultaneous interpreters also subject to the central processing bottleneck during language production, despite their professional experience?
2) If this is the case, is the stage they are subjected to the bottleneck shorter compared to bilinguals, and possibly similar to the performance of monolinguals when producing a word?
3) Can the suggested good anticipation skills of interpreters lead to more efficient lemma selection in comparison to non-trained bilinguals and monolinguals?

The following hypotheses are formulated:

1) This study expects that interpreters will be subject to the central processing bottleneck during lemma selection and phonological word-form selection, based on the results of existing research showing that monolinguals are subject to the bottleneck during language production.
2) Interpreters are expected to have a relatively shorter central processing bottleneck stage than bilinguals during word production. The expectation is based on previous research which demonstrates that extensive practice can reduce the central-processing bottleneck effect. Interpreters have extensive practice in language production under time pressure conditions. Although interpreters speak more than one language and need to fulfil language selection when producing a word, considering that professional interpreters are

experts in language with extensive experience in practice, a similar or slightly longer bottleneck stage is expected in the interpreter group compared to the monolinguals.

3) Monolinguals are expected to benefit more from medium constraint sentences than untrained bilinguals, which will facilitate the response time (RT) more in medium constraint sentences compared to low constraint sentences. This hypothesis is based on the evidence that bilinguals are worse than monolinguals at anticipating an upcoming word in sentence reading in L2 (Foucart et al., 2014). However, interpreters are suggested to be good at prediction (De Bot, 2000). Similar or slightly worse anticipation performance is expected among interpreters compared to monolinguals.

## 3.3 Experimental design

This study mimicked Experiment 1 of Ferreira and Pashler (2002), with three groups of respondents: professional simultaneous interpreters, untrained bilinguals and monolinguals. In the experiment, subjects conducted a dual task including a picture naming task (in context), and a non-linguistic sound discrimination task. Task 1 is a picture-naming task in sentence context. The sentences were presented visually instead of auditorily to avoid the possibility that any differences in results may be due to the faster processing of auditory information by interpreters than non-interpreters as a consequence of their working experience.

The sentences were shown one word per time at the centre of the screen, with the picture which required participants to name it as quickly as possible appearing at the end of each sentence. Two factors were manipulated to help explore the difference in lemma and phonological word-form selection: sentence constraint and word frequency. Cloze constraint, which is known to influence lemma selection (Butterworth, 1989; Roelofs, 1992; Levelt et al., 1999; Federmeier & Kutas, 2001), was manipulated. Medium-constraint sentences ease the selection of the picture name, while low-constraint sentences barely constrain the lemma selection of the following picture. The frequency of picture names was also manipulated, since frequency can influence phonological selection (Jescheniak & Levelt, 1994; Levelt & Wheeldon, 1994) (see Section 3.3.3 for more details). Task 2 was a tone discrimination task which included high, medium, and low pitch sounds. Three SOAs (50, 150, and 900 ms) were included to manipulate the

overlapping of the two tasks. A pilot study had been conducted before the main study. In the pilot study, PRP results were obtained among bilinguals and monolinguals during word production.

The design of this study was approved by the Macquarie University Faculty of Human Sciences Human Research Ethics Sub-Committee (5201600036) (see Appendix).

### 3.3.1 Sampling

The primary population for the study was the population of professional interpreters, and the sample was an availability sample from this group within Australia. For practical reasons the majority of the participants were from the Sydney region. The sampling of the bilingual and monolingual groups was done in a very deliberate way to match the participants in these groups with the participants sampled in the group of professional interpreters.

A total of ninety adults, resident in Australia, participated in the study. The participants consisted of three groups, each comprised of thirty participants: Group I consisted of professional simultaneous interpreters, Group II was comprised of proficient bilinguals, and Group III consisted of monolinguals. In Group II, bilinguals whose dominant language matched those of the interpreters were selected, while monolinguals (Group III) had English as their only language.[6] Professional simultaneous interpreters who have at least five years' working experience in simultaneous interpreting were recruited for Group I. The experience criterion was necessary to ensure that participants have had sufficient experience for the effects of professional practice to be evident. The languages of interpreters were: Chinese (8), Spanish (7), French (7), German (3), Japanese (3), Korean (1), and Portuguese (1). Proficient bilinguals who live or work in Australia participated in Group II. Their self-evaluated language proficiency in both languages is not statistically different from that of from Group I. Participants in Group II included translators (who do not do interpreting work), lecturers in universities located in Sydney, and students and post-doctoral students who use English as their working language. Main languages and dominant languages were matched between Group I and Group II, while age and gender were carefully matched for the three groups to make sure that there were no significant differences between the groups. For

[6] Some participants may have had very limited exposure to second-language learning in school, but they hardly ever use this knowledge, and it does not involve a fully developed or acquired language.

example, if the interpreter was a female, working as a French-English interpreter and vice versa, aged 60, also speaking Spanish, then her bilingual counterpart was a female whose most dominant languages were French and English, who might or not speak another/other language(s), 58-62 years old; and her monolingual counterpart was a female monolingual who can only speak English. The deviation of age range was within ± 2 years, expect for two subjects who were within ± 5 years, while one bilingual participant was within ± 10 years due to the difficulty of finding participants in the age range. The gender of one interpreter-monolingual pair was not matched and the dominant language of two interpreter-bilingual pairs, who speak the two same languages, was cross-matched due to the difficulty of finding participants.

Participants for Group I were recruited via email using the AIIC and NATTI websites[7] as well as through the networks of the participants and lecturers in SI in the Department of Linguistics at Macquarie University. Participants for Group II and Group III were recruited via email and notices on campus as well as through the social networks of participants. One interpreter was replaced because 40% of the trials of the whole experiment had to be discarded. Six bilinguals were replaced because more than 30% of the total trials had to be discarded, while four monolinguals were replaced in total because of the following reasons. One was replaced due to an error rate higher than 30%, one was replaced due to offering un-speeded responses (mean RT of Task 1 exceeded 3000 ms), and two monolinguals were replaced because they reported occasionally having short conversations in a second language.

### 3.3.2 Selection criteria

All participants were asked to fill out two questionnaires and take one test as part of the selection process. The questionnaires included: 1) the Edinburgh Handedness Inventory (MQ), used to measure the handedness of participants; 2) the Language Experience and Proficiency Questionnaire (LEAP-Q; Marian, Blumenfeld, & Kaushanskaya, 2007), used to measure participants' language background and experience. The test included one task: a semantic fluency task, used to measure the efficiency of word retrieval. Participants who

[7] The websites of the interpreting community. AIIC represents the International Association of Conference Interpreters, while NAATI represents the National Accreditation Authority for Translators and Interpreters.

were left-handed, or had a history of speech, language or hearing deficits,[8] or any other neurological deficits were excluded. Participants who were selected based on the outcomes from these questionnaires and test were called to participate in further experiments. No significant difference was obtained between the three groups on age range, educational background, and language background based on the questionnaires (see Table 3.1).

**Table 3. 1**

Interpreter, bilingual and monolingual participants' age and interpreters' simultaneous interpreting experience.

| | Interpreters | Bilinguals | Monolinguals |
|---|---|---|---|
| **Mean age** | **48.9** | **48.8** | **49.5** |
| **SI experience[9]** | **17.2** | | |
| **Education Years** | **20.0** | **18.1** | **17.9** |
| | **Proficiency of English** | | |
| **Speaking** | **9.0** | **8.5** | **9.6** |
| **Understanding** | **9.1** | **8.6** | **9.7** |
| **Reading** | **9.0** | **8.5** | **9.3** |
| | **Proficiency of other language** | | |
| **Speaking** | **9.5** | **9.4** | |
| **Understanding** | **9.6** | **9.5** | |
| **Reading** | **9.6** | **9.2** | |

**Note.** All the numbers were calculated by years

### 3.3.3 Apparatus and stimuli

The experiment mimicked Experiment 1 of Ferreira and Pashler (2002) by using the dual-task paradigm, and was conducted using the Neurobehavioral Systems—Presentation system with headset and Microsoft keyboard. Subjects were asked to conduct a picture-naming task (in context), and a non-linguistic sound discrimination task in each trial. Task 1 was a picture-naming task in rebus style. The materials are based on Experiment 2 in Griffin and Bock (1998), supplied directly via e-mail by Professor Griffin. Thus, all the materials were in English. The sentences were shown one word at a time at the centre of

[8] Considering the age range of the participants, this study allowed slight hearing loss to the extent that it did not interfere with performance in this study.
[9] The interpreting experience was specified as the number of years involved in simultaneous interpreting, specifically.

the screen, with the picture participants had to name as quickly as possible appearing at the end of each sentence. As outlined in section 3.3, two factors were manipulated: cloze constraint and word frequency. Cloze constraint is known to influence lemma selection (Butterworth, 1989; Roelofs, 1992; Levelt et al., 1999; Federmeier & Kutas, 2001). The frequency of picture names was also manipulated, since frequency can influence phonological selection (Jescheniak & Levelt, 1994; Levelt & Wheeldon, 1994).

Cloze constraint was manipulated by using low- and medium constraint sentences (high-constraint sentences were not used to avoid the ceiling effect). Medium-constraint sentences ease the selection of the picture name (e.g. the sentence "*The thief picked the…*(LOCK)"), while low-constraint sentences barely constrain the lemma selection of the following picture (e.g. "*At the end of the sentence was a drawing of a…*(LOCK)"). The frequency of the picture-word was also manipulated, using high-frequency and low-frequency names. Using the CELEX database (Baayen, Piepenbrock, & Van Rijn, 1993), high frequency words were defined as those occuring at a frequency of 100 per million words, while low frequency words were defined as those occurring at a frequency of 15 per million words. Task 2 was a tone discrimination task which included high, medium, and low pitch sounds. Three SOAs (50, 150, and 900 ms) were used to manipulate the overlapping of two tasks. The subject viewed the display from a distance of approximately 60 cm – 70 cm. The display was presented in bright white on a black background and was viewed under normal room illumination.

### 3.3.4 Design

The experiment was divided into 2 blocks (60 trials in each block) and the order of blocks was counterbalanced within groups.[10] Three independent variables were manipulated within each group: sentence constraint (medium vs. low), word frequency (high vs. low), and SOA (50, 150, and 900 ms). Cloze constraint, word frequency and SOA were manipulated within each subject. Word frequency was counterbalanced across blocks, while cloze constraint and SOA were counterbalanced within group. Each picture was presented once in each block, and none of the pictures was presented following the same

[10] The experiment was designed to have 2 blocks in order to guarantee that there would be enough trials under each condition to ensure statistically reliable results.

medium constraint sentence for each subject. The pictures presented under each experimental condition were rotated across subjects and each picture was presented 10 time under each experimental condition within a group (30 subjects). The sequence of trials was randomised, and the pitch tone was randomly selected.

### 3.3.5 Procedure

The subject was given instructions on the screen before each block. The instructions stated that the subject should make a verbal response to the picture and a button-push response to the pitch tone as quickly as possible. The importance of accuracy on both responses was emphasised, and participants were informed that the verbal response was recorded so that vocal errors could be detected. Subjects began with two practice blocks of 30 trials each. The first practice block was tone discrimination with 10 trials of each pitch. The second block was a dual-task practice block, with the same paradigm as the main blocks. None of the pictures and cloze sentences from the practice block were presented in the main block. During the dual-task practice block and main blocks, a sentence-repeated-section occurred between every fifteen or twenty trials, respectively. Subjects were asked to repeat the last sentence or provide the gist of the sentence showed on the screen to make sure they actually read the sentence. The participants then pressed the button to continue the trial after the sentence had been repeated.

Each trial began with a 500 ms blanked screen, followed by a plus sign as a fixation point which was displayed for 1000 ms at the centre of the screen (see Fig. 3.1). After the fixation point had disappeared, the foreperiod (500 ms) began. Then, the cloze sentence was presented by displaying each word in a rapid serial visual presentation (RSVP) paradigm in the centre of the screen. Each word lasted 285 ms and was shown in Times New Roman 12-point font. After the final word of the sentence had elapsed, the picture stimulus, which needed to be named as quickly as possible, was immediately presented and remained on the screen until two responses had been detected. The pre-determined SOA separating the picture stimulus and the tone discrimination task was 50, 150, or 900 ms. After the SOA had elapsed, a pitch tone (either 180 Hz, 500 Hz or 1200 Hz, defined as low, medium or high respectively) was presented for a duration of 285 ms. It was selected randomly and varied among trials, but each pitch of the tone was presented in equal numbers for each subject.

Subjects were required to respond to the tone pitch with their right hand, pressing the Red, Yellow, and Green colour keys for low, medium and high, respectively.

The feedback was displayed for 1500 ms, beginning after two responses had been detected. The following feedback options existed:

1. If the response order was correct, and a verbal and correct button-push response was given, the feedback message "Correct!" was given.
2. If the response order was correct, a verbal response was given, but an error was made on the button-push response, the feedback message was "Incorrect!"
3. If the order was wrong though a correct manual and verbal response had been given, a warning message was given: "You have the response in the wrong order!"
4. If a manual response had been given twice, then a warning message was showed: "Please only press button once!"
5. If a verbal response had been given twice, then a warning message was showed: "Please only respond verbally once!".

In the dual practice block, all the above five feedback options were given. In the main blocks, the "Correct!" feedback was excluded. The intertrial interval between the end of the previous trial (feedback or second response) and onset of the next trial (fixation point) was 1.3 s.

At the end of each block, the subject rested, and then continued by pressing the button when s\he felt ready. During this period, feedback was provided to the subject, focusing on the speed on both verbal and manual responses and the accuracy of the button-push response for the preceding block. Information on the total number of blocks as well as on how many block(s) the subject had finished was also shown before the next block instruction began.


**Figure 3. 1** The sequence of each trail in the main blocks.

## 3.4 Conclusion

This chapter has set out the research questions and hypotheses informing this study, and presented a detailed discussion of the experimental design. The following chapter presents the findings of the study in relation to the research questions and hypotheses, and discusses the conclusions that may be drawn from the findings.

# Chapter 4. Results and analysis

## 4.1 Introduction

In this chapter, the data of interpreter, bilingual and monolingual groups collected in the experiment are analysed and discussed. The results and analysis of the study are presented in Section 4.2, which is further divided into two sections: Section 4.2.1 presents the error rate analysis, and Section 4.2.2 sets out the RT analysis of the main task. The discussion of the analysed results is presented in Section 4.3, while Section 4.4 concludes the chapter with a brief summary.

In order to test the first hypothesis, namely that interpreters will be subject to the central processing bottleneck during lemma selection and phonological word-form selection, data was collected on the response time of the different groups. These data were also intended to test the second hypothesis, namely that interpreters would have a shorter central processing bottleneck than bilinguals during word production. Finally, in order to test the third hypothesis that monolinguals will benefit more from medium constraint sentences than untrained bilinguals, RT data was collected in medium and low constraint sentences. In addition to the RT, the error rate for responses was also measured.

## 4.2 Analysis

### 4.2.1 Error rate analysis

The reaction time and accuracies of each task were measured individually. For Task 1, any response latency under 200 ms or exceeding 2,500 ms were discarded as deviant (a total of 2.38%). Verbal responses which were the name of the picture but were not the intended names were discarded as well (for example, if the intended name of the picture was *scarf*, the response *muffler* was not the intended name and was discarded). While this may be a fruitful area of further investigation, in this study these errors needed to be excluded from the analysis, to avoid the influence of other possible factors on the duration of bottleneck and response latency. For example, the word frequency might be different (e.g. the

frequency of the words *scarf* and *muffler* is different), and the number of syllables, as well as the first phoneme of high and low frequency names may no longer be consistent. These kinds of responses were counted as incorrect verbal responses due to the difficulty of distinguishing whether the subjects were (suddenly) unable to access the intended word that this study expected and then chose its synonym (e.g. they forgot the word *scarf* and named the picture as *muffler*). The RT of Task 2 was not included in the analysis if its Task 1 trial was discarded. Trials on which the voice key failed to detect verbal responses, or the response order was reversed (a response to Task 2 was provided before responding to Task 1 first) were removed (a total of 2.59 % of trials). For Task 2, any response latencies faster than 200 ms or slower than 3,500 ms were discarded from the analysis. The error rates are presented in Table 4.1.

A multivariate analysis of variance was conducted to compare the performance between the three groups in these four error rates shown in Table 4.1: incorrect picture naming responses in Block 1, and Block 2; incorrect tone discrimination responses in Block 1, and Block 2. The significance level was .05 with 95% confidence-intervals. The results show that the number of errors for naming the pictures in Block 2 was significantly different between the three groups ($F$ (2, 87 = 7.517, $p$ = .001). The interpreter group made more errors than the bilingual and monolingual groups, while the monolinguals made the fewest errors. None of the other error rates reach significant differences between the three groups.

**Table 4.1**

Mean number of errors for Task 1 and Task 2 per subject for the interpreter group, bilingual group and monolingual group.

| **Conditions** | | **Interpreters** | **Bilinguals** | **Monolinguals** |
|---|---|---|---|---|
| | | **Picture naming errors** | | |
| **Block 1** | **ICR** | 4.5 | 3.8 | 3.55 |
| **Block 2** | **ICR** | 4.4 | 3.25 | 2.25 |
| | | **Tone discrimination errors** | | |
| **Block 1** | **ICR** | 2.75 | 4.7 | 5.25 |
| **Block 2** | **ICR** | 2.75 | 3.95 | 3.9 |

**Note.** "ICR" represents incorrect responses.

Interestingly, a significantly higher error rate was found for the interpreter group than for the bilinguals and monolinguals in the picture-naming task. However, as mentioned above, synonyms were also counted in the category of picture-naming errors with wrong picture names. Some interpreters appear to like providing more details even when instructed to name one simple word (e.g. they would name "eggs" as "broken eggs" or "eggs and bacon"). Thus, this error rate may not fully represent the true performance of the interpreters. This finding warrants further investigation.

### 4.2.2 Main task analysis

Task 1 and Task 2 were analysed with four-way 2×2×2×3 analyses of variance (ANOVAs), unless noted otherwise, with within-subjects variables of Blocks (Block 1 vs. Block 2), Constraints (medium constraint sentence vs. low constraint sentence), Frequencies (high frequency word vs. low frequency word), and SOAs (50, 150 or 900 ms), and between-subjects factors of Groups (interpreter group vs. bilingual group vs. monolingual group). The significance level for this analysis was .05 with 95% confidence-intervals.

#### 4.2.2.1 Task 1 analysis

The main effect of SOA was highly significant ($F$ (2, 174) = 63.086, $p$ < .001. MSe = 161,441.784), reflecting the slower RT at long SOA. Simple effect analyses show a significantly slower RT at long SOA than RT at short SOAs (197 ms slower than at SOA 150 ms and 214 ms slower than at 50 ms). The difference between the two shorter SOAs was only significant within the bilingual group but not for the interpreter and monolingual groups. These results reflect the "grouping" strategy discussed in Section 2.4.2.2, suggesting that prolonging RT1 at long SOA is a strategy of waiting for the response of Task 2. Therefore, the RTs at long SOA (900 ms) were discarded in Task 1 data analysis to guarantee that the response latency actually reflects the real performance of word production for the three groups. In other words, only the two shorter SOAs (50 ms and 150 ms) were included in the further Task 1 data analysis by using four-way 2×2×2×2 ANOVAs, with the same design mentioned above. Table 4.2 summarises all the effects, with F statistics, degrees of freedom and $p$-values associated with each factor.

**Table 4.2**

Analyses of variance results for Task 1: Picture naming task.

| Factors | Degrees of freedom | *F* value | *p* | |
|---|---|---|---|---|
| Group | 2, 87 | 4.170 | .019 | * |
| Blocks | 1, 87 | 174.626 | <.001 | *** |
| SOA | 1, 87 | 7.118 | .009 | ** |
| Constraints | 1, 87 | 19.772 | <.001 | *** |
| Frequency | 1, 87 | 106.036 | <.001 | *** |
| Constraints × Frequency | 1, 87 | 11.794 | .001 | ** |
| Blocks × Constraints | 1, 87 | 4.888 | .030 | * |
| Blocks × Frequency | 1, 87 | 24.608 | <.001 | *** |
| Blocks × Groups | 2, 87 | 5.312 | .007 | ** |
| Frequency × Groups | 2, 87 | 5.072 | .008 | ** |
| Constraints × Group | 2, 87 | .858 | .427 | |
| Constraints × SOA × Group | 2, 87 | 4.407 | .015 | * |
| Blocks × Frequency × SOA × Group | 2, 87 | 3.515 | .034 | * |

**Note.** * p < .05; ** p < .01; *** p < .001.

The **group difference** was statistically significant, showing that the bilinguals took a longer time in picture naming (1142 ms) than the interpreters (1061 ms), while the interpreters were slower than their monolingual counterparts (979 ms). The results of the post-hoc tests show that there was a significant difference between the bilingual and monolingual groups only, but none of other group pairs (see Fig. 4.1). The **main effect of blocks** was significant, and the RT in Block 2 is 178 ms faster than in Block 1. The **main effect of SOA** was significant. However, the mean RT at SOA 150 ms was only 17 ms slower than the mean RT at SOA 50 ms. The **constraint** and **frequency** effects were also significant, showing picture naming for medium-constraint sentences was significantly faster (24 ms) than for low-constraint sentences, while RT is significantly faster (74 ms) when word frequency is high than when it is low.

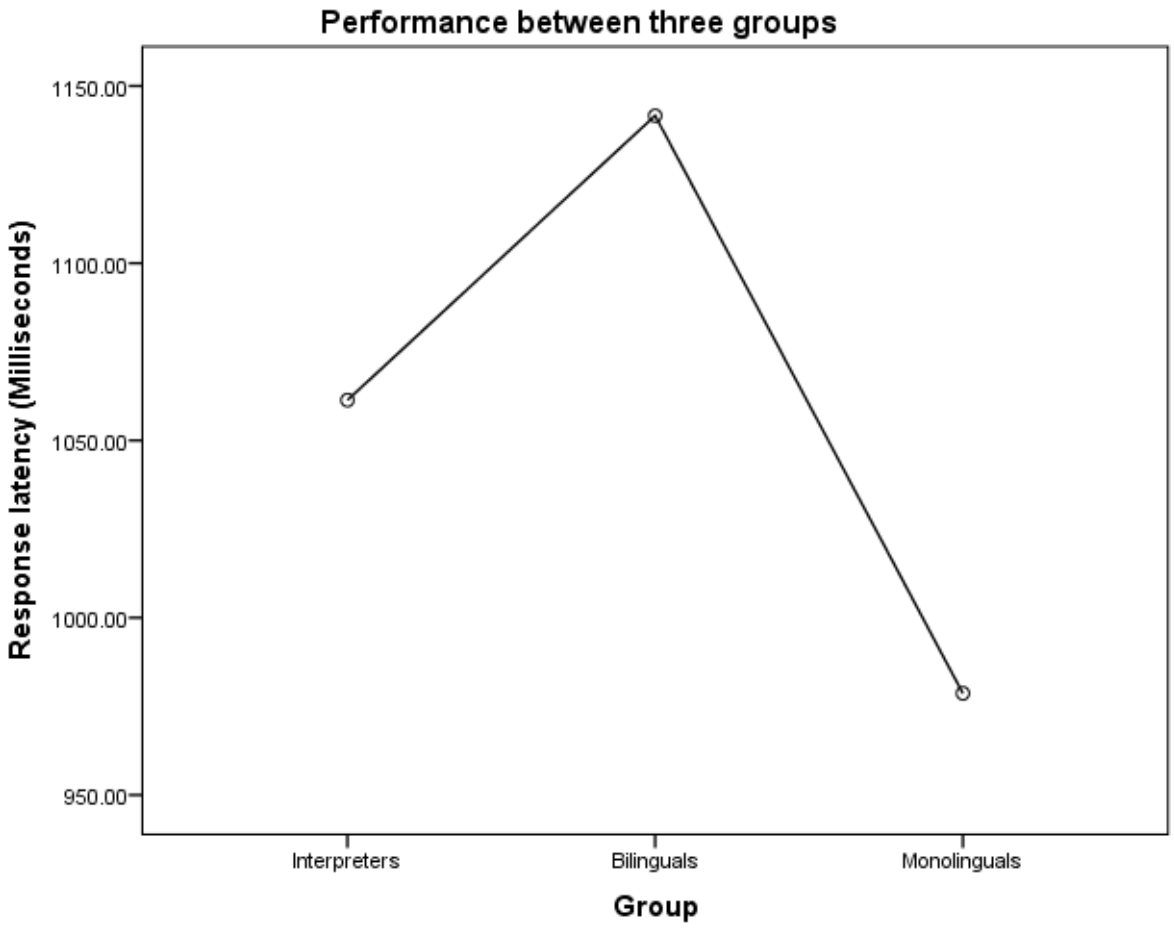



**Figure 4.1** The response latency of the picture naming among the three groups.

The **interaction of sentence constraint and word frequency** was significant. The result of the simple effects analysis shows that the RT in the medium-constraint sentence was only significantly faster than the RT in the low-constraint sentence when word frequency was low. The **interaction of blocks and sentence constraint** was significant. Simple effects analysis shows the difference in the picture-naming latency in the medium- and low-sentence constraint was only significant in Block 1 (38 ms difference) but not in Block 2 (11 ms difference). The **interaction of blocks and word frequency** was highly significant. The frequency effect was significant in both Block 1 and Block 2. However, the mean RT shows the high frequency words were named 99 ms faster than low frequency words in Block 1 but reduced to 49 ms in Block 2.

Four interactions with the group factor yield significant differences. The **interaction with blocks** was highly significant. In Block 1, simple effect analysis shows no statistically significant difference between the interpreter and bilingual groups (the interpreters were 73

ms faster than the bilinguals); a significant difference between the interpreter and monolingual groups (the interpreters were 132 ms slower than the monolinguals); and a significant difference between the bilingual and monolingual groups (the bilinguals were 206 ms slower than the monolinguals). In Block 2, however, only the performance of the bilinguals and monolinguals is significantly different, but none of the other group pairs.

The **interaction with word frequency** was also significant. Simple effect results show the word frequency effect was significant among all three groups. No significant difference was obtained between the interpreters and monolinguals, and interpreters and the bilinguals in the high and low word frequency condition. Highly significant differences between the bilinguals and monolinguals were sustained under both the word frequency conditions, showing bilinguals were 135 ms and 191 ms slower than monolinguals.

The three-way **interaction with sentence constraint and SOA** was significant. Simple effect analysis shows the constraint effect was significantly different among interpreters and bilinguals in the SOA 50 ms condition but not in the SOA 150 ms condition. The constraint difference was significant among the monolingual group at SOA 150 ms but not SOA 50 ms. The SOA effect was only significant in the medium constraint condition in the bilingual group. The RT at SOA 150 ms was 55 ms faster than the RT at SOA 50 ms.

**The four-way interaction between blocks, word frequency, SOA and groups** was significant. Simple effect analysis shows that the SOA effect reaches significance in bilinguals in Block 2 in the low word frequency condition. The word frequency effect was only not significant among the monolingual group at SOA 150 ms in Block 2.

The interaction between groups and constraint did not reach statistical significance. However, this study aimed to explore whether interpreters are as good at anticipation as monolinguals, and better than bilinguals; thus, simple effect analysis was conducted to further explore whether there is a difference between the three groups. The results show a significant constraint effect in the interpreter and monolingual groups, but not in the bilingual group (see Fig. 4.2). Interpreters benefit from the limited sentence information in the medium constraint condition, and named the picture 26 ms faster than in the low constraint condition. Monolinguals were 32 ms faster in picture naming in the medium

constraint condition than in the low constraint condition. However, bilinguals only showed a 15 ms difference in RT between the medium and low constraint sentence conditions.

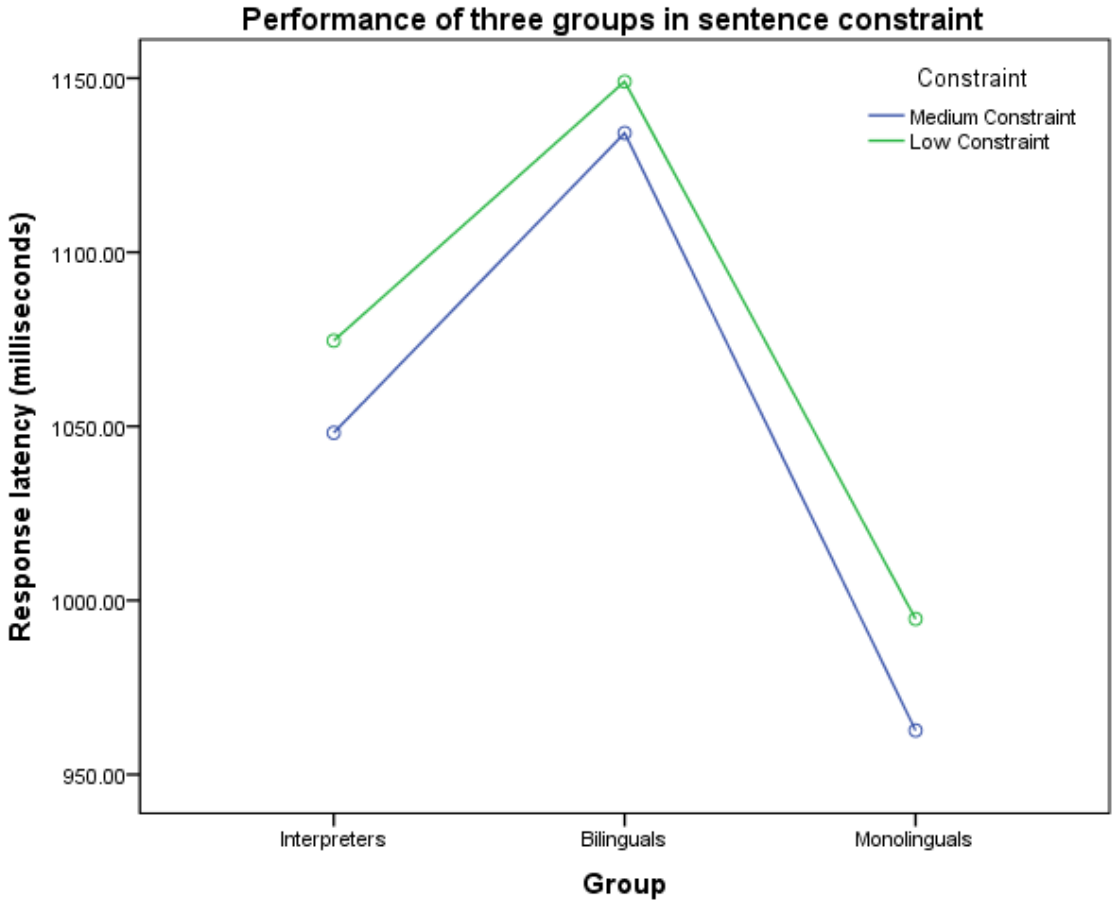



**Figure 4.2** The picture naming response latency in medium and low constraint sentences among the three groups.

### 4.2.2.2 Task 2 analysis

Table 4.3 shows the analyses of variance for Task 2. The main effect of **SOA** was highly significant, showing the longest RT (1630 ms) was obtained at short SOA (50 ms) while the shortest RT (1209 ms) was elicited at long SOA (900 ms). These results show that the PRP effect was obtained for all three groups in this study since the RT increased with reducing the SOA (see Fig 4.3). The main effect for **groups** was highly significant. Post-hoc results show that the difference between the interpreters and monolinguals was not significant. However, there were significant differences between monolinguals and bilinguals, and interpreters and bilinguals. The main effect of **block** was significant, showing the mean reaction time in Block 1 was statistically faster than in Block 2. The

main effects of **constraint** and **frequency** was highly significant. Responses were significantly faster (22 ms) in the medium sentence constraint condition than in the low sentence constraint condition, while the mean RT in the high frequency condition was significantly faster (73 ms) than in the low frequency condition. This result demonstrates that the manipulation of the first picture naming task in sentence constraint and word frequency propagate to the second tone discrimination task. This suggests that the stages impacted by sentence constraint and word frequency manipulations, namely, lemma and phonological word-form selection, are subject to the central processing bottleneck.

**Table 4.3**

Analyses of variance results for Task 2: Tone discrimination task.

| **Factors** | **Degrees of freedom** | ***F* value** | ***p*** | |
|---|---|---|---|---|
| SOA | 1, 174 | 343.769 | <.001 | *** |
| Group | 2, 87 | 8.063 | .001 | ** |
| Blocks | 1, 87 | 212.895 | <.001 | *** |
| Constraints | 1, 87 | 10.694 | .002 | ** |
| Frequency | 1, 87 | 74.497 | <.001 | *** |
| Constraints × Frequency | 1, 87 | 19.132 | <.001 | *** |
| Blocks × Constraints | 1, 87 | 11.949 | .001 | ** |
| Blocks × Frequency | 1, 87 | 4.103 | .046 | * |
| Blocks × SOA | 2, 174 | 3.506 | .032 | * |
| SOA × Groups | 4, 174 | 1.840 | .123 | |
| Blocks × Groups | 2, 87 | 4.748 | .011 | * |
| Frequency × Groups | 2, 87 | 4.659 | .012 | * |
| Blocks × Constraints × Frequency × Group | 2, 87 | 3.408 | .038 | * |

**Note.** * p < .05; ** p < .01; *** p < .001.

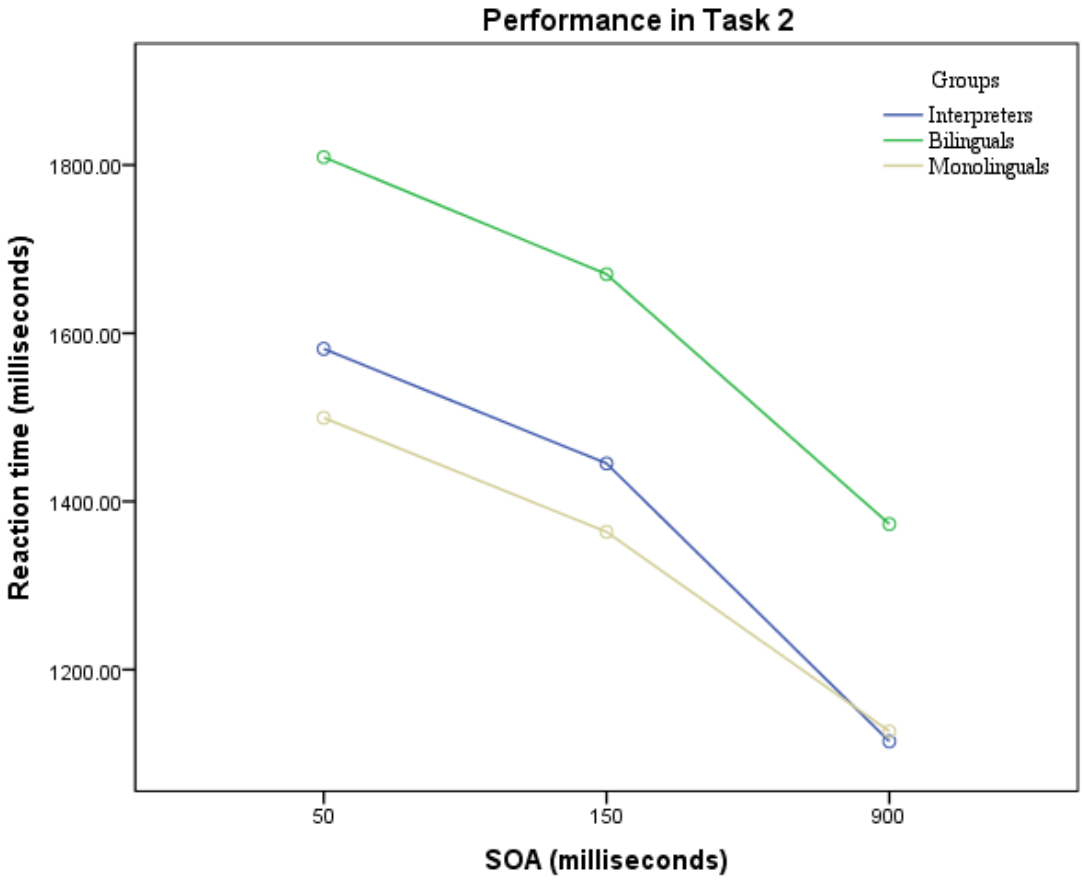



**Figure 4. 3** The performance of interpreters, bilinguals and monolinguals in Task 2.

The interaction between **blocks and SOA** was marginally significant. The analysis of simple effect results shows that the SOA effect was significant in both blocks, and the block effect was significant in all the SOAs. The difference between the effect for the short SOA and long SOA, however, was slightly smaller in Block 2 than in Block 1. However, the interaction between **groups and SOA** was not significant. Simple effect analysis shows a sustained difference between monolinguals and bilinguals, as well as interpreters and bilinguals across all SOAs, but not between monolinguals and interpreters. However, monolinguals show a shorter SOA difference (373 ms difference) than bilinguals (436 ms difference) and interpreters (467 ms difference).

In an attempt to further explore whether there is a difference between the three groups at the bottleneck stage during language production, three-way 2×2×2 ANOVAs were conducted with the factors of Blocks (Block 1 vs. Block 2), Constraints (medium constraint sentence vs. low constraint sentence) and Frequencies (high frequency word vs. low frequency word), to explore whether the RT difference between SOA 900 ms and

SOA 50 ms was similar between the three groups. The results show a marginally significant difference between the three groups at the bottleneck stage duration in word production ($F$ (2, 87) = 2.664; $p$ = .075). Post hoc results further show that interpreters have larger SOA difference than bilinguals and monolinguals, but the difference is only marginally significant between interpreters and monolinguals, but none of the other group pairs (see Fig 4.4). This result suggests that Task 2 was postponed more by the bottleneck, with lemma and phonological word-form selection being subject to the bottleneck in Task 1, and therefore, leading to larger SOA differences in the interpreter group than in the monolingual and bilingual groups. This result is inconsistent with the second hypothesis of this study, which postulated that interpreters would have a shorter bottleneck stage during word production than bilinguals as a consequence of SI experience.

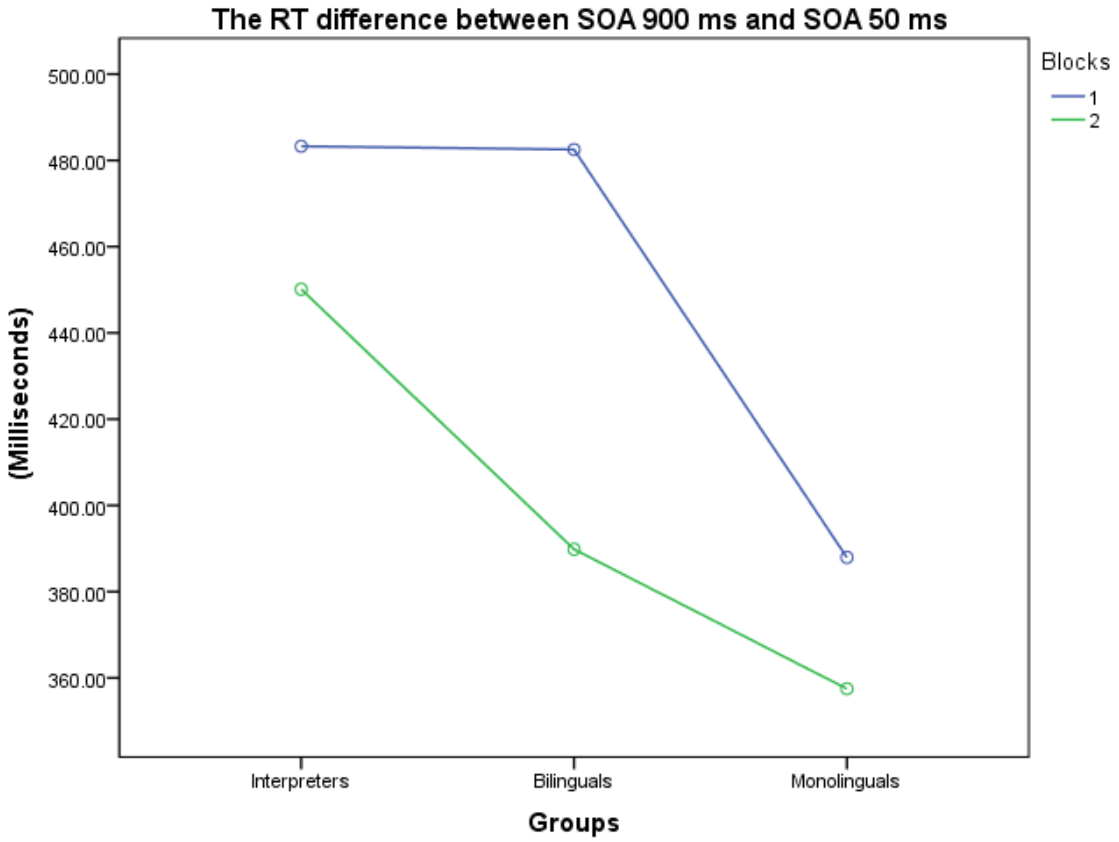



**Figure 4. 4** The performance of interpreter, bilingual and monolingual groups in the RT difference between SOA 900 and SOA 50 ms.

The **interaction between groups and blocks** was significant. The results of a simple effect analysis show that the difference between the bilingual group and the interpreter or monolingual group in each block was statistically significant. However, monolinguals have

a smaller block difference (160 ms difference) than bilinguals (256 ms difference) and interpreters (267 ms difference). The **interaction between blocks, sentence constraint, frequency and groups** was also significant. An analysis of the simple effect results shows that the performance of the bilingual group was significantly different from that of the interpreters and monolinguals in all conditions. However, there was no significant difference between the interpreters and the monolinguals except when naming the low frequency word pictures following the medium constraint sentences in Block 1. The sentence constraint effect was highly significant in the low frequency word condition in Block 1 in the bilingual and monolingual groups, and significant to some extent in the high frequency word condition in Block 2 in the monolingual group. The word frequency effect was not significant in the medium constraint condition in Block 1 in the bilingual and monolingual groups, and not significant in the medium constraint condition in Block 2 in the monolingual and interpreter groups.

## 4.3 Discussion

Consistent with the typical PRP effect, the reaction time to the second tone discrimination task increased with decreasing SOA, suggesting that word production is subject to the central processing bottleneck among monolinguals, bilinguals and even professional simultaneous interpreters (see Fig. 4.3), arguing against the assumption that all the tasks can be conducted in parallel by sharing the limited capacity, as has been argued by Schumacher et al. (2001), and specifically posited by some models of interpreting (Gile, 2009; Seeber, 2011). The sentence constraint effect and word frequency effect in Task 1 propagated to Task 2, consistent with the work of Ferreira and Pashler (2002), suggesting that the response selection stage in Task 2 cannot begin until the lemma and phonological word-form selection stages, on which sentence constraint and word frequency impact, respectively, are completed. In other words, lemma and phonological word-form selection in word production are subject to the central processing bottleneck, and therefore, postpone the response to the second unrelated task among all three groups.

Professional simultaneous interpreters are now known to also be subject to the central processing bottleneck during lemma and phonological word-form selection (confirming hypothesis 1), and might suffer the consequences of being subject to the bottleneck during

word production in SI. Contrary to hypothesis 2, the duration of the bottleneck stage during word production is not dramatically different among the three groups in this study (see Fig. 4.4). Monolinguals have the shortest bottleneck stage during word production while interpreters have the longest bottleneck stage, contrary to the prediction of hypothesis 2. However, this result might not actually reflect the performance between bilinguals and interpreters because the RT difference between long SOA and short SOA of bilinguals does not actually reflect the duration of their bottleneck stage. The word production latencies of bilinguals exceed 900 ms (see Fig. 4.1), in other words, exceed the long SOA. The bottleneck stage in Task 2 might still be postponed in the long SOA condition among the bilingual group. As a consequence, the RT difference between long and short SOA in this study might only reflect a smaller part of the bottleneck stage of bilinguals than of interpreters and monolinguals.

The RT difference between medium and low constraint sentences among interpreters and monolinguals yields a significant difference, suggesting that interpreters are as good at anticipation as monolinguals (confirming hypothesis 3). In contrast, for the bilingual group, the constraint effect did not reach significance, consistent with previous results which have shown that even proficient bilinguals cannot anticipate an upcoming word as well as monolinguals can (Martin, Thierry, Kuipers, Boutonnet, Foucart & Costa, 2013). What's more, interpreters appear to have developed particular strategies to ease the burden of conducting SI. The difference between interpreters and bilinguals was larger in Task 2 than in Task 1 (see Fig 4.1 and Fig 4.3), suggesting that the interpreters were more efficient in coordinating dual tasks than bilinguals. This result is consistent with research showing that professional simultaneous interpreters are better at processing dual tasks (Strobach, Becker, Schubert & Kühn, 2015), and furthermore, coordinating the bottleneck access (Sigman and Dehaene, 2006).

This study reiterates many of the findings of existing studies. In the first picture naming task, the naming latency was faster when the picture was followed by a medium constraint sentence than by a low constraint sentence. Pictures were named faster when the picture name was a high frequency word than when it was a low frequency word. Moreover, participants relied on the limited sentence information more when the presented picture name was a low frequency word than when it was high. These results are consistent with the work of Griffin and Bock (1998): pictures following medium constraint sentences are

named faster compared to low constraint sentences, while the pictures with high frequency names are named faster than low frequency words.

The performance of word production among the three groups seems quite different. The professional simultaneous interpreters appear to be faster in picture naming in context than their bilingual counterparts, while, however, still slower than their carefully matched monolingual counterparts, even if some interpreters and their bilingual counterparts have English as their dominant language. This result is in line with the results of Ivanova and Costa (2008), showing that proficient bilinguals have disadvantages in language production compared to monolinguals, even when testing their dominant and/or first language, supporting the assumption that bilinguals have disadvantages in linguistic tasks because of speaking more than one language (for more details, see Section 2.3.6, or Ivanova & Costa, 2008).

Although the pictures were presented in different constraint conditions in the two blocks, all three groups were still faster in picture naming when the same picture was presented again. The improved performance may not only be attributed to more efficient word retrieval and selection, but also possibly to the picture comprehension stage, since none of the pictures were presented to participants before the study. However, response times decreased less for monolinguals (110 ms faster in Block 2) than for interpreters (197 ms faster in Block 2) and bilinguals (222 ms faster in Block 2) when the pictures were presented for the second time. It may be that monolinguals were already quite fast in retrieving and producing a word. Thus, there was limited opportunity for the monolinguals to improve their performance compared to the other two groups. This might also be the case for the smaller word frequency difference for the monolingual group (48 ms difference) in comparison to the interpreter (74 ms difference) and bilingual (103 ms difference) groups, although the word frequency difference was much reduced in Block 2 for all three groups.

## 4.4 Conclusion

Simultaneous interpreters are shown to be subject to the central processing bottleneck during lemma and phonological word-form selection in word production, similar to

bilinguals and monolinguals. These results are consistent with hypothesis 1 of this study showing SI experience cannot eliminate the bottleneck stage during language production. Interpreters in this study are found to have a slightly longer bottleneck stage in language production than bilinguals and monolinguals, contrary to hypothesis 2 which inferred that interpreters will have a shorter bottleneck stage than bilinguals because of their extensive practice in language production. However, this hypothesis might not be completely ruled out since the difference between long SOA and short SOA in Task 2 might only reflect part of the bottleneck stage of bilinguals due to their long response latency in this study. Consistent with hypothesis 3, interpreters are good at anticipation and demonstrate a performance similar to their monolingual counterparts and better than their matched proficient bilinguals. It would therefore seem that interpreters are indeed able to benefit from their experience in countering the effects of the central bottleneck, rendering their word production closer to that of monolinguals than bilinguals.

# Chapter 5. Conclusion, limitations and avenues for further research

Following the work of Ferreira and Pashler (2002), which showed that language production has an impact on another concurrent non-linguistic task, this study explored whether professional simultaneous interpreters are also subject to the central processing bottleneck during word production. The study further aimed to explore whether SI experience, during which multiple concurrent tasks need to be conducted and speeded processing is required, helps interpreters reduce the bottleneck stage during language production. To do this, the study compared the performance of professional simultaneous interpreters with proficient bilinguals and monolinguals in the typical dual task paradigm which required speeded responses, and varied the interval of the two choice reaction time tasks to manipulate the overlapping of the bottleneck stage.

Challenging the assumption of capacity sharing models (Gile, 2009; Seeber, 2001), which suggest that all tasks can be conducted concurrently by sharing the limited capacity, this study has shown that simultaneous interpreters, who are known as time-sharing experts, are also subject to the central processing bottleneck when fulfilling lemma and phonological word-form selection during word production, just like bilinguals and monolinguals. That is, when fulfilling word production (more specifically, when fulfilling the selection of lemma and phonological word-form), another task, such as memory recall, has to wait until the word production has been completed (Rohrer & Pashler, 2003).

In other words, simultaneous interpreters encounter the bottleneck every time when they are producing a single word, and this bottleneck postpones another task and impairs their SI performance (below the level of conscious awareness). Therefore, there might be a negative correlation between SI performance and word production latency, as Christoffels and colleagues (2003) have suggested, such that the shorter production latency interpreters have, the shorter postponement the other task will be subject to, and the better SI performance interpreters might have. These findings demonstrate that language production is an important dimension of SI. As such it is worth undertaking further research on its

hidden impacts on SI, as well as considering this aspect in selecting and training interpreters.

The clear prolonged reaction time to the second non-linguistic task when reducing the interval between the two tasks, evident among interpreters, bilinguals and monolinguals in language production, not only suggests that the PRP effect is robust and ubiquitous, but also provides further support for Ferreira and Pashler's (2002) work, suggesting that monolinguals, bilinguals and even simultaneous interpreters are subject to the central processing bottleneck during lemma and phonological word-form selection.

Interpreters are found to be better at coordinating the dual task than bilinguals in this study, in line with the findings of Becker and colleagues (2016), whose work involves two non-linguistic tasks. These two studies therefore provide empirical support for the coordination feature in the Effort Model of Gile (2009), indicating that interpreters are cognitively skilled at coordinating multiple tasks, and furthermore, might be more efficient at planning the sequence of the bottleneck (Sigman & Dehaene, 2006) than untrained bilinguals, as a result of SI experience. Efficiency in coordinating multiple tasks might reduce the effort involved in delivering SI and raise its quality. Moreover, interpreters are shown to be as good at anticipation (known as an important quality of being a qualified interpreter) as monolinguals, and slightly better than their bilingual counterparts. Good anticipation helps efficient lemma selection and reduces the word production latency and duration that are subject to the bottleneck as a consequence.

The findings of this study should be interpreted against the background of certain limitations of the study. First, the definitions for the expertise and proficiency of the interpreters and bilinguals are inadequate. Years of interpreting experience, even SI experience, may not truly represent interpreting expertise. Similarly, determining the proficiency of the bilinguals in this study was problematic. Further refinements to the concepts of expertise and proficiency, as well as refinements to the ways in which these concepts are operationalised and controlled for in research of this kind, are needed. For example, for interpreters it is essential to have more fine-grained information about the frequency with which SI and CI are conducted. To control for expertise, assessments of the quality of output may be included. For bilinguals, more sophisticated tests of bilingual proficiency need to be explored.

Second, the long SOA is not long enough. As mentioned in section 4.3, the duration of the bottleneck stage during word production in this study might not fully represent the actual bottleneck duration of bilinguals. The picture naming latency of bilinguals far exceed 900 ms, and therefore, RT2 might still be postponed by the central stage in the first task even at long SOA condition. Such long RT results for the bilinguals were not expected and were not found in the pilot study, conceivably since most of the subjects (bilinguals and monolinguals) who participated in the pilot were around 20 to 30 years old whereas the mean age of participants in the experiment was just below 50. The bilinguals in the main study were recruited in accordance to the age, gender and languages of the interpreters who could be found and who were willing to participate in this study. Thus, the long SOA in this study was not long enough to compare the true duration of the bottleneck stage during word production among interpreters, bilinguals and monolinguals.

Third, experience in laboratory research may have played a role. The majority of the bilinguals and monolinguals who participated in this study were students, lecturers and even researchers. It is reasonable to assume that they may have had more experience in participating in a laboratory experiment, like this study, than the interpreters. As a result, they may have found the requirements of the study easier to handle, and consequently, have demonstrated better performance than interpreters. However, it should be pointed out that no such effects were observed among the interpreters, who followed the "rules" of the study and conducted it perfectly. Therefore, although this possibility might be a limitation of the study, no evident interference was observed.

Fourth, it should be pointed out that no intelligence testing was included in the set of tests conducted prior to the experiment. This is a potential limitation of the study. Intelligence is known to demonstrate a correlation with the duration of the response selection stage (Lee & Chabris, 2013), suggesting that people with high general cognitive ability experience a shorter bottleneck stage in non-linguistic tasks than people whose general cognitive ability is comparably lower. This study, however, mainly focused on linguistic tasks and explored the bottleneck stage in word production. Furthermore, no statistically significant difference was obtained between the three groups in educational background, suggesting that the general cognitive ability of three groups, logically, should not be much different. Furthermore, the materials used for testing in this study are all simple words and are

frequently encountered in daily life. That is, this study relied more on language proficiency than requiring efforts requiring a high degree of intelligent effort.

In sum, this study indicates that under the surface of ordinary "unimportant" language production, there actually exists a bottleneck which can postpone another task and might impair the performance of SI. These findings provide evidence against the capacity sharing model. This study also provides empirical support for three other important interpreting skills. First, interpreters are shown to be better at anticipation than proficient bilinguals. Second, interpreters are shown to be more efficient at coordinating dual tasks than proficient bilinguals. Third, interpreters are shown to be faster in lexical access than proficient bilinguals. This study suggests that more difficulties might be involved in SI than can be seen, and it is necessary to explore and unfold those mysteries to improve SI performance and SI training.

# References


Abutalebi, J., & Green, D. (2007). Bilingual language production: The neurocognition of language representation and control. *Journal of Neurolinguistics*, *20*(3), 242–275.

Allport, D. A., Antonis, B., & Reynolds, P. (1972). On the division of attention: A disproof of the single channel hypothesis. *The Quarterly Journal of Experimental Psychology*, *24*(2), 225–235.

Antón, E., Duñabeitia, J. A., Estévez, A., Hernández, J. A., Castillo, A., Fuentes, L.J., Davidson, D.J. and Carreiras, M. (2014). Is there a bilingual advantage in the ANT task? Evidence from children. *Frontiers in Psychology*, doi: 10.3389/fpsyg.2014.00398.

Baddeley, A. (2000). The episodic buffer: A new component of working memory? *Trends in Cognitive Sciences*, *4*(11), 417–423.

Baddeley, A. D., & Hitch, G. (1974). Working memory. *Psychology of Learning and Motivation*, *8*, 47–89.

Baddeley, A., Lewis, V., & Vallar, G. (1984). Exploring the articulatory loop. *The Quarterly Journal of Experimental Psychology*, *36*(2), 233–252.

Badecker, W., Miozzo, M., & Zanuttini, R. (1995). The two-stage model of lexical retrieval: Evidence from a case of anomia with selective preservation of grammatical gender. *Cognition*, *57*(2), 193–216.

Bakti, M. (2009). Speech disfluencies in simultaneous interpretation. In D. De Crom (Ed.), *Selected Papers of the CETRA Research Seminar in Translation Studies 2008*. https://www.arts.kuleuven.be/cetra/papers (retrieved 8 February 2017).

Barik, H. C. (1975). Simultaneous interpretation: Qualitative and linguistic data. *Language and Speech*, *18*(3), 272–297.

Berg, T. (1992). Productive and perceptual constraints on speech-error correction. *Psychological Research*, *54*(2), 114–126.

Berndt, R. S., Haendiges, A. N., Mitchum, C. C., & Sandson, J. (1997b). Verb retrieval in aphasia. 2. Relationship to sentence processing. *Brain and Language*, *56*(1), 107–137.

Berndt, R. S., Mitchum, C. C., Haendiges, A. N., & Sandson, J. (1997a). Verb retrieval in aphasia. 1. Characterizing single word impairments. *Brain and Language*, *56*(1), 68–106.

Bertelson, P. (1963). SR relationships and reaction times to new versus repeated signals in a serial task. *Journal of Experimental Psychology*, *65*(5), 478–484.

Bertelson, P. (1967). The time course of preparation. *Quarterly Journal of Experimental Psychology*, *19*(3), 272–279.

Bialystok, E. (2001). *Bilingualism in Development: Language, Literacy, and Cognition*. Cambridge: Cambridge University Press.

Bialystok, E. (2010). Bilingualism. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*(4), 559–572.

Bialystok, E., Craik, F. I., & Luk, G. (2012). Bilingualism: Consequences for mind and brain. *Trends in Cognitive Sciences*, *16*(4), 240–250.

Bialystok, E., Craik, F. I., Klein, R., & Viswanathan, M. (2004). Bilingualism, aging, and cognitive control: Evidence from the Simon task. *Psychology and Aging*, *19*(2), 290–303.

Bialystok, E., Craik, F., & Luk, G. (2008). Cognitive control and lexical access in younger and older bilinguals. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*(4), 859–873.

Biedermann, B., Ruh, N., Nickels, L., & Coltheart, M. (2008). Information retrieval in tip of the tongue states: New data and methodological advances. *Journal of Psycholinguistic Research*, *37*(3), 171–198.

Bloem, I., & La Heij, W. (2003). Semantic facilitation and semantic interference in word translation: Implications for models of lexical access in language production. *Journal of Memory and Language*, *48*(3), 468–488.

Bock, J. K. (1986). Syntactic persistence in language production. *Cognitive Psychology*, *18*(3), 355–387.

Bock, J. K., & Levelt, W. (1994). Language production: Grammatical encoding. In M. A. Gernsbacher (Ed.), *Handbook of Psycholinguistics*, pp. 945–984. San Diego, CA: Academic Press.

Boomer, D. S., & Laver, J. D. (1968). Slips of the tongue. *British Journal of Disorders of Communication*, *3*(1), 2–12.

Borger, R. (1963). The refractory period and serial choice-reactions. *Quarterly Journal of Experimental Psychology*, *15*(1), 1–12.

Broadbent, D. E. (2013). *Perception and Communication*. Oxford: Pergamon Press.

Brosseau-Lapré, F., & Rvachew, S. (2014). Cross-linguistic comparison of speech errors produced by English-and French-speaking preschool-age children with developmental

phonological disorders. *International Journal of Speech-Language Pathology*, *16*(2), 98–108.

Brown, A. S. (1991). A review of the tip-of-the-tongue experience. *Psychological Bulletin*, *109*(2), 204–223.

Burke, D. M., MacKay, D. G., Worthley, J. S., & Wade, E. (1991). On the tip of the tongue: What causes word finding failures in young and older adults? *Journal of Memory and Language*, *30*(5), 542–579.

Butterworth, B. (1981). Speech errors: Old data in search of new theories. *Linguistics*, *19*(19), 627–662.

Butterworth, B. (1989). Lexical access in speech production. In W. Marslen-Wilson (Ed.), *Lexical Representation and Process*, pp. 108–135. Cambridge, MA: MIT Press.

Caramazza, A. (1997). How many levels of processing are there in lexical access? *Cognitive Neuropsychology*, *14*(1), 177–208.

Caramazza, A., & Hillis, A. E. (1991). Lexical organization of nouns and verbs in the brain. *Nature*, *349*, 788–790.

Caramazza, A., & Miozzo, M. (1997). The relation between syntactic and phonological knowledge in lexical access: Evidence from the tip-of-the-tongue phenomenon. *Cognition*, *64*(3), 309–343.

Caramazza, A., & Miozzo, M. (1998). More is not always better: A response to Roelofs, Meyer, and Levelt. *Cognition*, *69*(2), 231–241.

Caramazza, A., Costa, A., Miozzo, M., & Bi, Y. (2001). The specific-word frequency effect: Implications for the representation of homophones in speech production. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*(6), 1430–1450.

Carrier, L. M., & Pashler, H. (1995). Attentional limits in memory retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*(5), 1339–1348.

Cenoz, J. (1998). *Pauses and Communication Strategies in Second Language Speech*. ERIC Document ED 426630. Rockville, MD: Educational Resources Information Center.

Chernov, G. V. (1979). Semantic aspects of psycholinguistic research in simultaneous interpretation. *Language and Speech*, *22*(3), 277–295.

Chernov, G. V. (2004). *Inference and Anticipation in Simultaneous Interpreting: A Probability-Prediction Model*. Amsterdam: John Benjamins.

Chincotta, D., & Underwood, G. (1998). Simultaneous interpreters and the effect of concurrent articulation on immediate memory: A bilingual digit span study. *Interpreting*, *3*(1), 1–20.

Christoffels, I. (2006). Listening while talking: The retention of prose under articulatory suppression in relation to simultaneous interpreting. *European Journal of Cognitive Psychology*, *18*(2), 206–220.

Christoffels, I. K., & de Groot, A. M. (2004). Components of simultaneous interpreting: Comparing interpreting with shadowing and paraphrasing. *Bilingualism: Language and Cognition*, *7*(3), 227–240.

Christoffels, I. K., De Groot, A. M., & Kroll, J. F. (2006). Memory and language skills in simultaneous interpreters: The role of expertise and language proficiency. *Journal of Memory and Language*, *54*(3), 324–345.

Christoffels, I. K., De Groot, A. M., & Waldorp, L. J. (2003). Basic skills in a complex task: A graphical model relating memory and lexical retrieval to simultaneous interpreting. *Bilingualism: Language and Cognition*, *6*(3), 201–211.

Christoffels, I. K., Firk, C., & Schiller, N. O. (2007). Bilingual language control: An event-related brain potential study. *Brain Research*, *1147*, 192–208.

Colomé, À. (2001). Lexical activation in bilinguals' speech production: Language-specific or language-independent? *Journal of Memory and Language*, *45*(4), 721–736.

Cook, A. E., & Meyer, A. S. (2008). Capacity demands of phoneme selection in word production: New evidence from dual-task experiments. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*(4), 886–899.

Costa, A., & Caramazza, A. (1999). Is lexical selection in bilingual speech production language-specific? Further evidence from Spanish–English and English–Spanish bilinguals. *Bilingualism: Language and Cognition*, *2*(3), 231–244.

Costa, A., & Santesteban, M. (2004). Lexical access in bilingual speech production: Evidence from language switching in highly proficient bilinguals and L2 learners. *Journal of Memory and Language*, *50*(4), 491–511.

Costa, A., Alario, F. X., & Caramazza, A. (2005). On the categorical nature of the semantic interference effect in the picture-word interference paradigm. *Psychonomic Bulletin & Review*, *12*(1), 125–131.

Costa, A., Caramazza, A., & Sebastian-Galles, N. (2000). The cognate facilitation effect: Implications for models of lexical access. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(5), 1283–1296.

Costa, A., Colomé, À., & Caramazza, A. (2000). Lexical access in speech production: The bilingual case. *Psicológica*, *21*(2), 403–437.

Costa, A., Mahon, B., Savova, V., & Caramazza, A. (2003). Level of categorisation effect: A novel effect in the picture-word interference paradigm. *Language and Cognitive Processes*, *18*(2), 205–234.

Costa, A., Miozzo, M., & Caramazza, A. (1999). Lexical selection in bilinguals: Do words in the bilingual's two lexicons compete for selection? *Journal of Memory and Language*, *41*(3), 365–397.

Costa, A., Roelstraete, B., & Hartsuiker, R. J. (2006). The lexical bias effect in bilingual speech production: Evidence for feedback between lexical and sublexical levels across languages. *Psychonomic Bulletin & Review*, *13*(6), 972–977.

Costa, A., Santesteban, M., & Ivanova, I. (2006). How do highly proficient bilinguals control their lexicalization process? Inhibitory and language-specific selection mechanisms are both functional. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*(5), 1057–1074.

Coughlan, A. K., & Warrington, E. K. (1981). The impairment of verbal semantic memory: A single case study. *Journal of Neurology, Neurosurgery & Psychiatry*, *44*(12), 1079–1083.

Cutting, J. C., & Ferreira, V. S. (1999). Semantic and phonological information flow in the production lexicon. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*(2), 318–344.

Damasio, A. R., & Tranel, D. (1993). Nouns and verbs are retrieved with differently distributed neural systems. *Proceedings of the National Academy of Sciences*, *90*(11), 4957–4960.

Damian, M. F., & Martin, R. C. (1999). Semantic and phonological codes interact in single word production. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*(2), 345–361.

Daniele, A., Giustolisi, L., Silveri, M. C., Colosimo, C., & Gainotti, G. (1994). Evidence for a possible neuroanatomical basis for lexical processing of nouns and verbs. *Neuropsychologia*, *32*(11), 1325–1341.

Darò, V. (1994). Non-linguistic factors influencing simultaneous interpretation. In S. Lambert & B. Moser-Mercer (Eds.), *Bridging the Gap: Empirical Research in Simultaneous Interpretation*, pp. 249–269. Amsterdam: John Benjamins.

Darò, V., & Fabbro, F. (1994). Verbal memory during simultaneous interpretation: Effects of phonological interference. *Applied Linguistics*, *15*(4), 365–381.

De Abreu, P. M. E., Cruz-Santos, A., Tourinho, C. J., Martin, R., & Bialystok, E. (2012). Bilingualism enriches the poor enhanced cognitive control in low-income minority children. *Psychological Science*, *23*(11), 1364–1371.

De Bot, K. (1992). A bilingual production model: Levelt's 'speaking' model adapted. *Applied Linguistics*, *13*(1), 1–24.

De Bot, K. (2000). Simultaneous interpreting as language production. In B. E. Dimitrova & K. Hyltenstam (Eds.), *Language Processing and Simultaneous Interpreting: Interdisciplinary Perspectives*, pp. 65–88. Amsterdam: John Benjamins.

De Bot, K. (2004). The multilingual lexicon: Modelling selection and control. *International Journal of Multilingualism*, *1*(1), 17–32.

De Groot, A. M. (1992). Determinants of word translation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*(5), 1001–1018.

De Groot, A. M. (1997). The cognitive study of translation and interpretation: Three approaches. In J. H. Danks, G. M. Shreve, S. B. Fountain & M. K. McBeath (Eds.), *Cognitive Processes in Translation and Interpreting*, pp. 25–56. Thousand Oaks, Calif: Sage Publications.

De Groot, A. M. (2000). A complex-skill approach to translation. In S. Tirkkonen-Condit & R. Jääskeläinen (Eds.), *Tapping and Mapping the Processes of Translation and Interpreting: Outlooks on Empirical Research*, pp. 53–68. Amsterdam: John Benjamins.

De Groot, A. M., & Nas, G. L. (1991). Lexical representation of cognates and noncognates in compound bilinguals. *Journal of Memory and Language*, *30*(1), 90–123.

De Jong, R. (1993). Multiple bottlenecks in overlapping task performance. *Journal of Experimental Psychology: Human Perception and Performance*, *19*(5), 965–980.

De Renzi, E., & Di Pellegrino, G. (1995). Sparing of verbs and preserved, but ineffectual reading in a patient with impaired word production. *Cortex*, *31*(4), 619–636.

Declerck, M., & Kormos, J. (2012). The effect of dual task demands and proficiency on second language speech production. *Bilingualism: Language and Cognition*, 15(4), 782–796.

Declerck, M., & Philipp, A. M. (2015). A review of control processes and their locus in language switching. *Psychonomic Bulletin & Review*, *22*(6), 1630–1645.

Defrancq, B. (2015). Corpus-based research into the presumed effects of short EVS. *Interpreting*, 17(1), 26–45.

Dell, G. S. (1985). Positive feedback in hierarchical connectionist models: Applications to language production. *Cognitive Science*, *9*(1), 3–23.

Dell, G. S. (1986). A spreading-activation theory of retrieval in sentence production. *Psychological Review*, *93*(3), 283.

Dell, G. S. (1988). The retrieval of phonological forms in production: Tests of predictions from a connectionist model. *Journal of Memory and Language*, *27*(2), 124–142.

Dell, G. S. (1990). Effects of frequency and vocabulary type on phonological speech errors. *Language and Cognitive Processes*, *5*(4), 313–349.

Dell, G. S., & O'Seaghdha, P. G. (1991). Mediated and convergent lexical priming in language production: A comment on Levelt et al. (1991). *Psychological Review*, *98*(4), 604–614.

Dell, G. S., & O'Seaghdha, P. G. (1992). Stages of lexical access in language production. *Cognition*, *42*(1), 287–314.

Dell, G. S., & Reich, P. A. (1981). Stages in sentence production: An analysis of speech error data. *Journal of Verbal Learning and Verbal Behavior*, *20*(6), 611–629.

Dell, G. S., Schwartz, M. F., Martin, N., Saffran, E. M., & Gagnon, D. A. (1997). Lexical access in aphasic and nonaphasic speakers. *Psychological Review*, *104*(4), 801–838.

Deutsch, J. A., & Deutsch, D. (1963). Attention: Some theoretical considerations. *Psychological Review*, *70*(1), 80–90.

Dijkstra, T., & Van Heuven, W. J. (1998). The BIA model and bilingual word recognition. In J. Grainger & A. M. Jacobs (Eds.), *Localist Connectionist Approaches to Human Cognition*, pp. 189–225. Hove: Psychology Press.

Dimond, S. J. (1970). Hemispheric refractoriness and control of reaction time. *The Quarterly Journal of Experimental Psychology*, *22*(4), 610–617.

Donnelly, S. (2016). *Re-Examining the Bilingual Advantage on Interference-Control and Task-Switching Tasks: A Meta-Analysis*. CUNY Academic Works.

Fabbro, F., & Gran, L. (1994). Neurological and neuropsychological aspects of polyglossia and simultaneous interpretation. In S. Lambert & B. Moser-Mercer (Eds.), *Bridging the Gap: Empirical Research in Simultaneous Interpretation*, pp. 273–317. Amsterdam: John Benjamins.

Fabbro, F., Gran, B., & Gran, L. (1991). Hemispheric specialization for semantic and syntactic components of language in simultaneous interpreters. *Brain and Language*, *41*(1), 1–42.

Fay, D., & Cutler, A. (1977). Malapropisms and the structure of the mental lexicon. *Linguistic Inquiry*, *8*(3), 505–520.

Federmeier, K. D., & Kutas, M. (2001). Meaning and modality: Influences of context, semantic memory organization, and perceptual predictability on picture processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*(1), 202–224.

Ferreira, V. S., & Pashler, H. (2002). Central bottleneck influences on the processing stages of word production. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(6), 1187–1199.

Finkbeiner, M., & Caramazza, A. (2006). Now you see it, now you don't: On turning semantic interference into facilitation in a Stroop-like task. *Cortex*, *42*(6), 790–796.

Finkbeiner, M., Almeida, J., Janssen, N., & Caramazza, A. (2006). Lexical selection in bilingual speech production does not involve language suppression. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*(5), 1075–1089.

Finkbeiner, M., Gollan, T., & Caramazza, A. (2006). Bilingual lexical access: What is the (hard) problem? *Bilingualism: Language and Cognition*, *9*, 153–166.

Foygel, D., & Dell, G. S. (2000). Models of impaired lexical access in speech production. *Journal of Memory and Language*, *43*(2), 182–216.

Friedman, A., Polson, M. C., & Dafoe, C. G. (1988). Dividing attention between the hands and the head: Performance trade-offs between rapid finger tapping and verbal memory. *Journal of Experimental Psychology: Human Perception and Performance*, *14*(1), 60–68.

Fromkin, V. (1973). *Slips of the Tongue*. San Francisco: WH Freeman.

Fromkin, V. A. (1971). The non-anomalous nature of anomalous utterances. *Language*, 27–52.

Garrett, M. F. (1975). The analysis of sentence production. *Psychology of Learning and Motivation*, *9*, 133–177.

Garrett, M. F. (1976). Syntactic processes in sentence production. *New Approaches to Language Mechanisms*, *30*, 231–256.

Garrett, M. F. (1980). Levels of processing in sentence production. *Language Production*, *1*, 177–220.

Garrett, M. F. (1982). Production of speech: Observations from normal and pathological language use. *Normality and Pathology in Cognitive Functions*, 19–76.

Garrett, M. F. (1988). Processes in language production. In J. F. Newmeyer (Ed.), *Linguistics: The Cambridge Survey* (Vol. 3), pp. 69–96. Cambridge: Cambridge University Press.

Gerver, D. (1974a). The effects of noise on the performance of simultaneous interpreters: Accuracy of performance. *Acta Psychologica*, *38*(3), 159–167.

Gerver, D. (1974b). Simultaneous listening and speaking and retention of prose. *The Quarterly Journal of Experimental Psychology*, *26*(3), 337–341.

Gerver, D. (1975). A psychological approach to simultaneous interpretation. *Meta: Journal des Traducteurs / Meta:Translators' Journal*, *20*(2), 119–128.

Gerver, D. (1976). Empirical studies of simultaneous interpretation: A review and a model. In R. W. Briskin (Ed.), *Translation: Applications and Research*, pp. 165–207. New York: Gardner Press.

Gile, D. (1997). Conference interpreting as a cognitive management problem. In J. H. Danks, G. M. Shreve, S. B. Fountain, & M. K. McBeath (Eds.), *Cognitive Processes in Translation and Interpreting*, pp. 196–214. Thousand Oaks, CA: Sage.

Gile, D. (2009). *Basic Concepts and Models for Interpreter and Translator Training*. Amsterdam: John Benjamins.

Glaser, W. R., & Düngelhoff, F. J. (1984). The time course of picture-word interference. *Journal of Experimental Psychology: Human Perception and Performance*, *10*(5), 640–654.

Gold, B. T., Kim, C., Johnson, N. F., Kryscio, R. J., & Smith, C. D. (2013). Lifelong bilingualism maintains neural efficiency for cognitive control in aging. *The Journal of Neuroscience*, *33*, 387–396.

Gollan, T. H., & Silverberg, N. B. (2001). Tip-of-the-tongue states in Hebrew–English bilinguals. *Bilingualism: Language and Cognition*, *4*(1), 63–83.

Gollan, T. H., Ferreira, V. S., Cera, C., & Flett, S. (2014). Translation-priming effects on tip-of-the-tongue states. *Language, Cognition and Neuroscience*, *29*(3), 274–288.

Gollan, T. H., Montoya, R. I., Fennema-Notestine, C., & Morris, S. K. (2005). Bilingualism affects picture naming but not picture classification. *Memory & Cognition*, *33*(7), 1220–1234.

Gonon, M. H., Bruckert, R., & Michel, F. (1989). Lexicalization in an anomic patient. *Neuropsychologia*, *27*(4), 391–407.

Goodglass, H., Kaplan, E., Weintraub, S., & Ackerman, N. (1976). The "tip-of-the-tongue" phenomenon in aphasia. *Cortex*, *12*(2), 145–153.

Green, D. W. (1986). Control, activation, and resource: A framework and a model for the control of speech in bilinguals. *Brain and Language*, *27*(2), 210–223.

Green, D. W. (1993). Towards a model of L2 comprehension and production. In R. Schreuder & B. Weltens (Eds.), *The Bilingual Lexicon*, pp. 249–277. Amsterdam: John Benjamins.

Green, D. W. (1998). Mental control of the bilingual lexico-semantic system. *Bilingualism: Language and Cognition*, *1*(2), 67–81.

Griffin, Z. M., & Bock, K. (1998). Constraint, word frequency, and the relationship between lexical processing levels in spoken word production. *Journal of Memory and Language*, *38*(3), 313–338.

Griffin, Z. M., & Ferreira, V. S. (2006). Properties of spoken language production. In M. Traxler & M. A. Gemsbacher (Eds.), *Handbook of Psycholinguistics*, pp. 21–59. London: Elsevier.

Grosjean, F. (2001). The bilingual's language modes. In J. L. Nicol (Ed.), *One Mind, Two Languages: Bilingual Language Processing*, pp. 1– 22. Oxford, U.K.: Blackwell.

Guo, T., Liu, F., Chen, B., & Li, S. (2013). Inhibition of non-target languages in multilingual word production: Evidence from Uighur–Chinese–English trilinguals. *Acta Psychologica*, *143*(3), 277–283.

Harley, T. A. (1984). A critique of top-down independent levels models of speech production: Evidence from non-plan-internal speech errors. *Cognitive Science*, *8*(3), 191–219.

Hazeltine, E., Teague, D., & Ivry, R. B. (2002). Simultaneous dual-task performance reveals parallel response selection after practice. *Journal of Experimental Psychology: Human Perception and Performance*, *28*(3), 527–545.

Hermans, D., Bongaerts, T., De Bot, K., & Schreuder, R. (1998). Producing words in a foreign language: Can speakers prevent interference from their first language? *Bilingualism: Language and Cognition*, *1*, 213–229.

Hervais-Adelman, A. G., Moser-Mercer, B., & Golestani, N. (2011). Executive control of language in the bilingual brain: Integrating the evidence from neuroimaging to neuropsychology. *Frontiers in Psychology, 2:234*, doi: 10.3389/fpsyg.2011.00234.

Hilchey, M. D., & Klein, R. M. (2011). Are there bilingual advantages on nonlinguistic interference tasks? Implications for the plasticity of executive control processes. *Psychonomic Bulletin & Review*, *18*(4), 625–658.

Hillis, A. E., & Caramazza, A. (1991). Mechanisms for accessing lexical representations for output: Evidence from a category-specific semantic deficit. *Brain and Language*, *40*(1), 106–144.

Hillis, A. E., & Caramazza, A. (1995). Representation of grammatical categories of words in the brain. *Journal of Cognitive Neuroscience*, *7*(3), 396–407.

Hillis, A. E., Rapp, B., Romani, C., & Caramazza, A. (1990). Selective impairment of semantics in lexical processing. *Cognitive Neuropsychology*, *7*(3), 191–243.

Hiltunen, S., Pääkkönen, R., Vik, G. V., & Krause, C. M. (2014). On interpreters' working memory and executive control. *International Journal of Bilingualism*. *20*(3),297–314.

Ibáñez, A. J., Macizo, P., & Bajo, M. T. (2010). Language access and language selection in professional translators. *Acta Psychologica*, *135*(2), 257–266.

Injoque-Ricle, I., Barreyro, J. P., Formoso, J., & Jaichenco, V. I. (2015). Expertise, working memory and articulatory suppression effect: Their relation with simultaneous interpreting performance. *Advances in Cognitive Psychology*, *11*(2), 56–63.

Isham, W. P. (1994). Memory for sentence form after simultaneous interpretation: Evidence both for and against deverbalization. In S. Lambert & B. Moser-Mercer (Eds.), *Bridging the Gap: Empirical Research in Simultaneous Interpretation*, pp. 191–211. Amsterdam: John Benjamins.

Ivanova, I., & Costa, A. (2008). Does bilingualism hamper lexical access in speech production? *Acta Psychologica*, *127*(2), 277–288.

Ivry, R. B., Franz, E. A., Kingstone, A., & Johnston, J. C. (1998). The psychological refractory period effect following callosotomy: Uncoupling of lateralized response codes. *Journal of Experimental Psychology: Human Perception and Performance*, *24*(2), 463–480.

Jackson, G. M., Swainson, R., Cunnington, R., & Jackson, S. R. (2001). ERP correlates of executive control during repeated language switching. *Bilingualism: Language and Cognition*, *4*(2), 169–178.

Janssen, N., Schirm, W., Mahon, B. Z., & Caramazza, A. (2008). Semantic interference in a delayed naming task: Evidence for the response exclusion hypothesis. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*(1), 249–256.

Jared, D., Poh, R. P. Y., & Paivio, A. (2013). L1 and L2 picture naming in Mandarin–English bilinguals: A test of bilingual dual coding theory. *Bilingualism: Language and Cognition*, *16*(2), 383–396.

Jescheniak, J. D., & Levelt, W. J. (1994). Word frequency effects in speech production: Retrieval of syntactic information and of phonological form. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(4), 824–843.

Jescheniak, J. D., & Schriefers, H. (1998). Discrete serial versus cascaded processing in lexical access in speech production: Further evidence from the coactivation of near-synonyms. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*(5), 1256–1274.

Jescheniak, J. D., Meyer, A. S., & Levelt, W. J. (2003). Specific-word frequency is not all that counts in speech production: Comments on Caramazza, Costa, et al. (2001) and new experimental data. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*(6), 1430–1450.

Johnsen, A. M., & Briggs, G. E. (1973). On the locus of display load effects in choice reactions. *Journal of Experimental Psychology*, *99*(2), 266–271.

Kahneman, D. (1973). *Attention and Effort*. Englewood Cliffs, NJ: Prentice-Hall.

Katz, J. J., & Fodor, J. A. (1963). The structure of a semantic theory. *Language*, *39*(2), 170–210.

Kay, J., & Ellis, A. (1987). A cognitive neuropsychological case study of anomia. *Brain*, *110*(3), 613–629.

Keele, S. W. (1973). *Attention and Human Performance*. Pacific Palisades, CA: Goodyear.

Kempen, G., & Huijbers, P. (1983). The lexicalization process in sentence production and naming: Indirect election of words. *Cognition*, *14*(2), 185–209.

Kikyo, H., Ohki, K., & Sekihara, K. (2001). Temporal characterization of memory retrieval processes: An fMRI study of the 'tip of the tongue' phenomenon. *European Journal of Neuroscience*, *14*(5), 887–892.

Kirchhoff, H. (1976). Simultaneous interpreting: Interdependence of variables in the interpreting process, interpreting models and interpreting strategies. In F.Pöchhacker & M. Shlesinger (Eds), *The Interpreting Studies Reader*, pp. 111–119. London: Routledge.

Kleinman, D. (2013). Resolving semantic interference during word production requires central attention. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(6), 1–32, doi:10.1037/a0033095.

Kleiss, J. A., & Lane, D. M. (1986). Locus and persistence of capacity limitations in visual information processing. *Journal of Experimental Psychology: Human Perception and Performance*, *12*(2), 200–210.

Köpke, B., & Nespoulous, J. L. (2006). Working memory performance in expert and novice interpreters. *Interpreting*, 8(1), 1–23.

Köpke, B., & Signorelli, T. M. (2012). Methodological aspects of working memory assessment in simultaneous interpreters. *International Journal of Bilingualism*, *16*(2), 183–197.

Kousaie, S., & Phillips, N. A. (2012). Ageing and bilingualism: Absence of a "bilingual advantage" in Stroop interference in a nonimmigrant sample. *The Quarterly Journal of Experimental Psychology*, *65*(2), 356–369.

Kremin, H., & Basso, A. (1993). Apropos the mental lexicon: The naming of nouns and verbs. In F. J. Stachowiak, R. De Bleser, G. Deloche, R. Kaschel, H. Kremin, P. North, L. Pizzamiglio, I. Robertson & B. A. Wilson A. (Eds.), *Developments in the Assessment and Rehabilitation of Brain Damaged Patients*. Tubingen: Gunter Narr Verlag.

Kroll, J. F., & Stewart, E. (1994). Category interference in translation and picture naming: Evidence for asymmetric connections between bilingual memory representations. *Journal of Memory and Language*, *33*(2), 149–174.

Kroll, J. F., Bobb, S. C., Misra, M., & Guo, T. (2008). Language selection in bilingual speech: Evidence for inhibitory processes. *Acta Psychologica*, *128*(3), 416–430.

La Heij, W. (1988). Components of Stroop-like interference in picture naming. *Memory & Cognition*, *16*(5), 400–410.

La Heij, W. (2005). Selection processes in monolingual and bilingual lexical access. In J. F. Kroll & A. M. B. de Groot (Eds.), *Handbook of Bilingualism: Psycholinguistic Approaches*, pp. 289–307. New York: Oxford University Press.

La Heij, W., Hooglander, A., Kerling, R., & Van Der Velden, E. (1996). Nonverbal context effects in forward and backward word translation: Evidence for concept mediation. *Journal of Memory and Language*, *35*(5), 648–665.

La Heij, W., Mak, P., Sander, J., & Willeboordse, E. (1998). The gender-congruency effect in picture-word tasks. *Psychological Research*, *61*(3), 209–219.

Lachman, R. (1973). Uncertainty effects on time to access the internal lexicon. *Journal of Experimental Psychology*, *99*(2), 199–208.

Lambert, S. (1988). Information processing among conference interpreters: A test of the depth-of-processing hypothesis. *Meta: Journal des Traducteurs Meta:/Translators' Journal*, *33*(3), 377–387.

Le Dorze, G., & Nespoulous, J. L. (1989). Anomia in moderate aphasia: Problems in accessing the lexical representation. *Brain and Language*, *37*(3), 381–400.

Lee, J. J., & Chabris, C. F. (2013). General cognitive ability and the psychological refractory period: Individual differences in the mind's bottleneck. *Psychological Science*. *24*(7), 1226–1233.

Levelt, W. J. (1993). *Speaking: From Intention to Articulation.* Cambridge, MA: MIT Press.

Levelt, W. J. (1999). Models of word production. *Trends in Cognitive Sciences, 3*(6), 223–232.

Levelt, W. J., & Wheeldon, L. (1994). Do speakers have access to a mental syllabary? *Cognition*, *50* (1-3), 239–269.

Levelt, W. J., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences*, *22*(1), 1–38.

Levelt, W. J., Schriefers, H., Vorberg, D., Meyer, A. S., Pechmann, T., & Havinga, J. (1991). The time course of lexical access in speech production: A study of picture naming. *Psychological Review*, *98*(1), 122–142.

Levy, J., & Pashler, H. (2001). Is dual-task slowing instruction dependent? *Journal of Experimental Psychology: Human Perception and Performance*, *27*(4), 862–869.

Levy, J., Pashler, H., & Boer, E. (2006). Central interference in driving: Is there any stopping the psychological refractory period? *Psychological Science*, *17*(3), 228–235.

Liepelt, R., & Prinz, W. (2011). How two share two tasks: Evidence for a social PRP effect. *Experimental Brain Research*, *221*, 387–396.

Lijewska, A., & Chmiel, A. (2015). Cognate facilitation in sentence context–translation production by interpreting trainees and non-interpreting trilinguals. *International Journal of Multilingualism*, *12*(3), 358–375.

Linck, J. A., Schwieter, J. W., & Sunderman, G. (2012). Inhibitory control predicts language switching performance in trilingual speech production. *Bilingualism: Language and Cognition*, *15*(3), 651–662.

Liu, M. (2008). How do experts interpret? Implications from research in interpreting studies and cognitive science. In G. Hansen, A. Chesterman & H. Gerzymisch-Arbogast (Eds.), *Efforts and Models in Interpreting and Translation Research: A Tribute to Daniel Gile* (Vol. 80), pp. 159–178. Amsterdam: John Benjamins.

Liu, M., Schallert, D. L., & Carroll, P. J. (2004). Working memory and expertise in simultaneous interpreting. *Interpreting*, *6*(1), 19–42.

Logan, G. D., & Burkell, J. (1986). Dependence and independence in responding to double stimulation: A comparison of stop, change, and dual-task paradigms. *Journal of Experimental Psychology: Human Perception and Performance*, *12*(4), 549–563.

Luk, G., Bialystok, E., Craik, F. I., & Grady, C. L. (2011). Lifelong bilingualism maintains white matter integrity in older adults. *Journal of Neuroscience*, 31 (46), 16808–16813.

Lupker, S. J. (1979). The semantic nature of response competition in the picture-word interference task. *Memory & Cognition*, *7*(6), 485–495.

Macnamara, B. N., & Conway, A. R. (2014). Novel evidence in support of the bilingual advantage: Influences of task demands and experience on cognitive control and working memory. *Psychonomic Bulletin & Review*, *21*(2), 520–525.

Macnamara, J., & Kushnir, S. L. (1971). Linguistic independence of bilinguals: The input switch. *Journal of Verbal Learning and Verbal Behavior*, *10*(5), 480–487.

Mädebach, A., Oppermann, F., Hantsch, A., Curda, C., & Jescheniak, J. D. (2011). Is there semantic interference in delayed naming? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 37(2), 522–538.

Mahon, B. Z., Costa, A., Peterson, R., Vargas, K. A., & Caramazza, A. (2007). Lexical selection is not by competition: A reinterpretation of semantic interference and facilitation effects in the picture-word interference paradigm. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*(3), 503–535.

Marian, V., Blumenfeld, H. K., & Kaushanskaya, M. (2007). The Language Experience and Proficiency Questionnaire (LEAP-Q): Assessing language profiles in bilinguals and multilinguals. *Journal of Speech, Language, and Hearing Research*, *50*(4), 940–967.

Martin, C. D., Thierry, G., Kuipers, J. R., Boutonnet, B., Foucart, A., & Costa, A. (2013). Bilinguals reading in their second language do not predict upcoming words as native readers do. *Journal of Memory and Language*, *69*(4), 574–588.

Martin-Rhee, M. M., & Bialystok, E. (2008). The development of two types of inhibitory control in monolingual and bilingual children. *Bilingualism: Language and Cognition*, *11*(01), 81–93.

McCarthy, R., & Warrington, E. K. (1985). Category specificity in an agrammatic patient: The relative impairment of verb retrieval and comprehension. *Neuropsychologia*, *23*(6), 709–727.

McLeod, P. (1977). Parallel processing and the psychological refractory period. *Acta Psychologica*, *41*(5), 381–396.

Meuter, R. F., & Allport, A. (1999). Bilingual language switching in naming: Asymmetrical costs of language selection. *Journal of Memory and Language*, *40*(1), 25–40.

Meyer, A. S., & Schriefers, H. (1991). Phonological facilitation in picture-word interference experiments: Effects of stimulus onset asynchrony and types of interfering stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *17*(6), 1146–1160.

Meyer, D. E., & Kieras, D. E. (1997). A computational theory of executive cognitive processes and multiple-task performance: Part I. Basic mechanisms. *Psychological Review*, *104*(1), 3–65.

Meyer, D. E., Kieras, D. E., Lauber, E., Schumacher, E. H., Glass, J., Zurbriggen, E., Gmeindl, L., & Apfelblat, D. (1995). Adaptive executive control: Flexible multiple-task performance without pervasive immutable response-selection bottlenecks. *Acta Psychologica*, *90*, 163–190.

Miceli, G., Silveri, M. C., Nocentini, U., & Caramazza, A. (1988). Patterns of dissociation in comprehension and production of nouns and verbs. *Aphasiology*, *2*, 351–358.

Miceli, G., Silveri, M. C., Villa, G., & Caramazza, A. (1984). On the basis for the agrammatic's difficulty in producing main verbs. *Cortex*, *20*(2), 207–220.

Miozzo, M., & Caramazza, A. (1997). Retrieval of lexical-syntactic features in tip-of-the tongue states. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*(6), 1410–1423.

Mor, B., Yitzhaki-Amsalem, S., & Prior, A. (2014). The joint effect of bilingualism and ADHD on executive functions. *Journal of Attention Disorders: A Journal of Theoretical and Applied Science, 19*(6), 527–541.

Morales, J., Padilla, F., Gómez-Ariza, C. J., & Bajo, M. T. (2015). Simultaneous interpretation selectively influences working memory and attentional networks. *Acta Psychologica*, *155*, 82–91.

Moser, B. (1978). Simultaneous interpretation: A hypothetical model and its practical application. In D. Gerver & H. W. Sinaiko (Eds.), *Language Interpretation and Communication,* pp. 353–368. New York: Plenum Press.

Navon, D., & Miller, J. (2002). Queuing or sharing? A critical evaluation of the single-bottleneck notion. *Cognitive Psychology*, *44*(3), 193–251.

Neisser, U., Novick, R., & Lazar, R. (1963). Searching for ten targets simultaneously. *Perceptual and Motor Skills*, *17*(3), 955–961.

Nickerson, R. S. (1965). Response time to the second of two successive signals as a function of absolute and relative duration of intersignal interval. *Perceptual and Motor Skills*, *21*(1), 3–10.

Norman, D. A., & Shallice, T. (1986). Attention to action: Willed and automatic control of behavior. In R. J. Davidson, G. E. Schwartz & D. Shapiro (Eds.), *Consciousness and Self-regulation*, pp. 1–18. New York: Plenum.

Oldfield, R. C., & Wingfield, A. (1965). Response latencies in naming objects. *Quarterly Journal of Experimental Psychology*, *17*(4), 273–281.

Osgood, C. E. (1963). On understanding and creating sentences. *American Psychologist*, *18*(12), 735–751.

Osman, A., & Moore, C. M. (1993). The locus of dual-task interference: Psychological refractory effects on movement-related brain potentials. *Journal of Experimental Psychology: Human Perception and Performance*, *19*(6), 1292–1312.

Paap, K. R., & Greenberg, Z. I. (2013). There is no coherent evidence for a bilingual advantage in executive processing. *Cognitive Psychology*, *66*(2), 232–258.

Padilla, F., Bajo, M. T., & Macizo, P. (2005). Articulatory suppression in language interpretation: Working memory capacity, dual tasking and word knowledge. *Bilingualism: Language and Cognition*, *8*(3), 207–219.

Paivio, A. (1971). Imagery and deep structure in the recall of English nominalizations. *Journal of Verbal Learning and Verbal Behavior*, *10*(1), 1–12.

Paradis, M. (1994). Toward a neurolinguistic theory of simultaneous translation: The framework. *International Journal of Psycholinguistics*, *10*(3), 319–335.

Paradis, M., Goldblum, M. C., & Abidi, R. (1982). Alternate antagonism with paradoxical translation behavior in two bilingual aphasic patients. *Brain and Language*, *15*(1), 55–69.

Pashler, H & Christian, C. L. (1994). Bottlenecks in planning and producing vocal, manual, and foot responses. Center for Human Information Processing Technical Report, University of California at San Diego, La Jolla, California.

Pashler, H. (1984). Processing stages in overlapping tasks: Evidence for a central bottleneck. *Journal of Experimental Psychology: Human Perception and Performance*, *10*(3), 358–377.

Pashler, H. (1989). Dissociations and dependencies between speed and accuracy: Evidence for a two-component theory of divided attention in simple tasks. *Cognitive Psychology*, *21*(4), 469–514.

Pashler, H. (1990). Do response modality effects support multiprocessor models of divided attention? *Journal of Experimental Psychology: Human Perception and Performance*, *16*(4), 826–842.

Pashler, H. (1993). Doing two things at the same time. *American Scientist*, *81*(1), 48–55.

Pashler, H. (1994). Dual-task interference in simple tasks: Data and theory. *Psychological Bulletin*, *116*(2), 220–244.

Pashler, H., & Baylis, G. C. (1991). Procedural learning: II. Intertrial repetition effects in speeded-choice tasks. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *17*(1), 33–48.

Pashler, H., & Johnston, J. C. (1989). Chronometric evidence for central postponement in temporally overlapping tasks. *The Quarterly Journal of Experimental Psychology*, *41*(1), 19–45.

Pashler, H., & O'Brien, S. (1993). Dual-task interference and the cerebral hemispheres. *Journal of Experimental Psychology: Human Perception and Performance*, *19*(2), 315–330.

Pashler, H., Carrier, M., & Hoffman, J. (1993). Saccadic eye movements and dual-task interference. *The Quarterly Journal of Experimental Psychology*, *46*(1), 51–82.

Pelham, S. D., & Abrams, L. (2014). Cognitive advantages and disadvantages in early and late bilinguals. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(2), 313–325.

Peterson, R. R., & Savoy, P. (1998). Lexical selection and phonological encoding during language production: Evidence for cascaded processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*(3), 539–557.

Poarch, G. J., & Van Hell, J. G. (2012). Executive functions and inhibitory control in multilingual children: Evidence from second-language learners, bilinguals, and trilinguals. *Journal of Experimental Child Psychology*, *113*(4), 535–551.

Pöchhacker, F. (2016). *Introducing Interpreting Studies*. London: Routledge.

Poulisse, N. (1997). Language production in bilinguals. In A. M. B. De Groot & J. F. Kroll (Eds.), *Tutorials in Bilingualism: Psycholinguistic Perspectives*, pp. 201–224. Mahwah, NJ: Erlbaum

Poulisse, N. (1999). *Slips of the Tongue: Speech Errors in First and Second Language Production*. Amsterdam: John Benjamins.

Poulisse, N., & Bongaerts, T. (1994). First language use in second language production. *Applied Linguistics*, *15*(1), 36–57.

Rapp, B., & Caramazza, A. (1997). The modality-specific organization of grammatical categories: Evidence from impaired spoken and written sentence production. *Brain and Language*, *56*(2), 248–286.

Roelofs, A. (1992). A spreading-activation theory of lemma retrieval in speaking. *Cognition*, *42*(1-3), 107–142.

Roelofs, A. (1993). Testing a non-decompositional theory of lemma retrieval in speaking: Retrieval of verbs. *Cognition*, *47*(1), 59–87.

Roelofs, A. (1997). The WEAVER model of word-form encoding in speech production. *Cognition*, *64*(3), 249–284.

Rohrer, D., & Pashler, H. E. (2003). Concurrent task effects on memory retrieval. *Psychonomic Bulletin & Review*, *10*(1), 96–103.

Rohrer, D., Pashler, H., & Etchegaray, J. (1998). When two memories can and cannot be retrieved concurrently. *Memory & Cognition*, *26*(4), 731–739.

Rohrman, N. L. (1970). More on the recall of nominalizations. *Journal of Verbal Learning and Verbal Behavior*, *9*(5), 534–536.

Ruthruff, E., Miller, J., & Lachmann, T. (1995). Does mental rotation require central mechanisms? *Journal of Experimental Psychology: Human Perception and Performance*, *21*(3), 552–570.

Ruthruff, E., Pashler, H. E., & Hazeltine, E. (2003). Dual-task interference with equal task emphasis: Graded capacity sharing or central postponement? *Perception & Psychophysics*, *65*(5), 801–816.

Ruthruff, E., Pashler, H. E., & Klaassen, A. (2001). Processing bottlenecks in dual-task performance: Structural limitation or strategic postponement? *Psychonomic Bulletin & Review*, *8*(1), 73–80.

Schiller, N. O., & Caramazza, A. (2002). The selection of grammatical features in word production: The case of plural nouns in German. *Brain and Language*, *81*(1), 342–357.

Schriefers, H. (1993). Syntactic processes in the production of noun phrases. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*(4), 841–850.

Schriefers, H., & Jescheniak, J. D. (1999). Representation and processing of grammatical gender in language production: A review. *Journal of Psycholinguistic Research*, *28*(6), 575–600.

Schriefers, H., Meyer, A. S., & Levelt, W. J. (1990). Exploring the time course of lexical access in language production: Picture-word interference studies. *Journal of Memory and Language*, *29*(1), 86–102.

Schroeder, S. R., & Marian, V. (2012). A bilingual advantage for episodic memory in older adults. *Journal of Cognitive Psychology*, *24*(5), 591–601.

Schubert, T. (1999). Processing differences between simple and choice reactions affect bottleneck localization in overlapping tasks. *Journal of Experimental Psychology: Human Perception and Performance, 25*(2), 408–425.

Schumacher, E. H., Seymour, T. L., Glass, J. M., Fencsik, D. E., Lauber, E. J., Kieras, D. E., & Meyer, D. E. (2001). Virtually perfect time sharing in dual-task performance: Uncorking the central cognitive bottleneck. *Psychological Science, 12*(2), 101–108.

Schweda-Nicholson, N. (1987). Linguistic and extralinguistic aspects of simultaneous interpretation. *Applied Linguistics*, *8*(2), 194–205.

Seeber, K. G. (2011). Cognitive load in simultaneous interpreting: Existing theories—new models. *Interpreting*, *13*(2), 176–204.

Seeber, K. G. (2001). Intonation and anticipation in simultaneous interpreting. *Cahiers de Linguistique Française*, *23*, 61–97.

Segal, S. J., & Fusella, V. (1970). Influence of imaged pictures and sounds on detection of visual and auditory signals. *Journal of Experimental Psychology*, *83*(3), 458–464.

Shatzman, K. B., & Schiller, N. O. (2004). The word frequency effect in picture naming: Contrasting two hypotheses using homonym pictures. *Brain and Language*, *90*(1), 160–169.

Sigman, M., & Dehaene, S. (2006). Dynamics of the central bottleneck: Dual-task and task uncertainty. *PLoS Biology*, *4*(7), 1227–1238.

Signorelli, T. M., Haarmann, H. J., & Obler, L. K. (2011). Working memory in simultaneous interpreters: Effects of task and age. *International Journal of Bilingualism, 16*(2) 198–212.

Spalek, K., & Schriefers, H. J. (2005). Dominance affects determiner selection in language production. *Journal of Memory and Language*, *52*(1), 103–119.

Starreveld, P. A. (2000). On the interpretation of onsets of auditory context effects in word production. *Journal of Memory and Language*, *42*(4), 497–525.

Starreveld, P. A., & La Heij, W. (1995). Semantic interference, orthographic facilitation, and their interaction in naming tasks. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*(3), 686–698.

Starreveld, P. A., & La Heij, W. (1996). Time-course analysis of semantic and orthographic context effects in picture naming. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*(4), 896–918.

Starreveld, P. A., De Groot, A. M., Rossmark, B. M., & Van Hell, J. G. (2014). Parallel language activation during word processing in bilinguals: Evidence from word production in sentence context. *Bilingualism: Language and Cognition*, *17*(2), 258–276.

Stemberger, J. P. (1985). An interactive activation model of language production. *Progress in the Psychology of Language*, *1*, 143–186.

Stemberger, J. P. (1985). Bound morpheme loss errors in normal and agrammatic speech: One mechanism or two? *Brain and Language*, *25*(2), 246–256.

Strobach, T., Becker, M., Schubert, T., & Kühn, S. (2015). Better dual-task processing in simultaneous interpreters. *Frontiers in Psychology*, *6*, doi:10.3389/fpsyg.2015.01590.

Strobach, T., Schütz, A., & Schubert, T. (2015). On the importance of Task 1 and error performance measures in PRP dual-task studies. *Frontiers in Psychology*, *6*, doi: 10.3389/fpsyg.2015.00403.

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, *18*(6), 643–662.

Tao, L., Marzecová, A., Taft, M., Asanowicz, D., & Wodniecka, Z. (2011). The efficiency of attentional networks in early and late bilinguals: The role of age of acquisition. *Frontiers in Psychology*, *2*, 83–99.

Telford, C. W. (1931). The refractory phase of voluntary and associative responses. *Journal of Experimental Psychology*, *14*(1), 1–36.

Tombu, M., & Jolicœur, P. (2003). A central capacity sharing model of dual-task performance. *Journal of Experimental Psychology: Human Perception and Performance*, *29*(1), 3–18.

Treisman, A. M. (1964). Verbal cues, language, and meaning in selective attention. *American Journal of Psychology*, 77, 206–219.

Treisman, A. M., & Davies, A. (1973). Divided attention to ear and eye. In S. Kornblum (Ed.), *Attention and Performance IV*, pp. 101–117. San Diego, CA: Academic Press.

Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, *12*(1), 97–136.

Treisman, A. M., & Riley, J. G. (1969). Is selective attention selective perception or selective response? A further test. *Journal of Experimental Psychology*, *79*, 27–34.

Van Besien, F. (1999). Anticipation in simultaneous interpretation. *Meta: Journal des Traducteurs*, *44*(2), 250–259.

Van Hell, J. G., & De Groot, A. M. (1998a). Conceptual representation in bilingual memory: Effects of concreteness and cognate status in word association. *Bilingualism: Language and Cognition*, *1*(3), 193–211.

Van Hell J. G., & De Groot A. M. B. (1998b). Disentangling context availability and concreteness in lexical decision and word translation. *The Quarterly Journal of Experimental Psychology: Section A*, *51*(1), 41–63.

Van Hell, J. G., & Dijkstra, T. (2002). Foreign language knowledge can influence native language performance in exclusively native contexts. *Psychonomic Bulletin & Review*, *9*(4), 780–789.

Vigliocco, G., Antonini, T., & Garrett, M. F. (1997). Grammatical gender is on the tip of Italian tongues. *Psychological Science*, *8*(4), 314–317.

Vince, M. A. (1948). The intermittency of control movements and the psychological refractory period. *British Journal of Psychology General Section*, *38*(3), 149–157.

Wang, B., & Li, T. (2015). An empirical study of pauses in Chinese-English simultaneous interpreting. *Perspectives*, *23*(1), 124–142.

Welford, A. T. (1952). The 'psychological refractory period' and the timing of high-speed performance: A review and a theory. *British Journal of Psychology General Section*, *43*(1), 2–19.

Wellman, H. M. (1977). Tip of the tongue and feeling of knowing experiences: A developmental study of memory monitoring. *Child Development*, *48*, 13–21.

Wickens, C. D. (1980). The structure of attentional resources. In R. S. Nickerson (Ed.), *Attention and Performance VIII*, pp. 239–257. Hillsdale, NJ: Erlbaum.

Yang, S., Yang, H., & Lust, B. (2011). Early childhood bilingualism leads to advances in executive attention: Dissociating culture and language. *Bilingualism: Language and Cognition*, *14*(3), 412–422.

Yudes, C., Macizo, P., & Bajo, T. (2012). Coordinating comprehension and production in simultaneous interpreters: Evidence from the articulatory suppression effect. *Bilingualism: Language and Cognition*, *15*(02), 329–339.

Yudes, C., Macizo, P., Morales, L., & Bajo, M. T. (2013). Comprehension and error monitoring in simultaneous interpreters. *Applied Psycholinguistics*, *34*(5), 1039–1057.

Zingeser, L. B., & Berndt, R. S. (1988). Grammatical class and context effects in a case of pure anomia: Implications for models of language production. *Cognitive Neuropsychology*, *5*(4), 473–516.

# Appendix

**LONGJIAO SUI <longjiao.sui@students.mq.edu.au>**

## RE: HS Ethics Application - Approved (5201600036)(Con/Met)

1 message

**Fhs Ethics** <fhs.ethics@mq.edu.au> 11 March 2016 at 14:25

To: Dr Haidee Kruger <haidee.kruger@mq.edu.au>

Cc: Ms Helen Slatyer <helen.slatyer@mq.edu.au>, Associate Professor Jan-Louis Kruger <janlouis.kruger@mq.edu.au>, Miss Longjiao Sui <longjiao.sui@students.mq.edu.au>

Dear Dr Kruger,

Re: "Are simultaneous interpreters subject to the central processing bottleneck during language production?"(5201600036)

Thank you very much for your response. Your response has addressed the issues raised by the Faculty of Human Sciences Human Research Ethics Sub-Committee and approval has been granted, effective 11th March 2016. This email constitutes ethical approval only.

This research meets the requirements of the National Statement on Ethical Conduct in Human Research (2007). The National Statement is available at the following web site:

http://www.nhmrc.gov.au/_files_nhmrc/publications/attachments/e72.pdf.

The following personnel are authorised to conduct this research:

Associate Professor Jan-Louis Kruger
Dr Haidee Kruger
Miss Longjiao Sui
Ms Helen Slatyer

Please note the following standard requirements of approval:

1. The approval of this project is conditional upon your continuing

compliance with the National Statement on Ethical Conduct in Human Research (2007).

2. Approval will be for a period of five (5) years subject to the provision of annual reports.

Progress Report 1 Due: 11th March 2017
Progress Report 2 Due: 11th March 2018
Progress Report 3 Due: 11th March 2019
Progress Report 4 Due: 11th March 2020
Final Report Due: 11th March 2021

NB. If you complete the work earlier than you had planned you must submit a Final Report as soon as the work is completed. If the project has been discontinued or not commenced for any reason, you are also required to submit a Final Report for the project.

Progress reports and Final Reports are available at the following website:

http://www.research.mq.edu.au/current_research_staff/human_research_ethics/application_resources

3. If the project has run for more than five (5) years you cannot renew approval for the project. You will need to complete and submit a Final Report and submit a new application for the project. (The five year limit on renewal of approvals allows the Sub-Committee to fully re-review research in an environment where legislation, guidelines and requirements are continually changing, for example, new child protection and privacy laws).

4. All amendments to the project must be reviewed and approved by the Sub-Committee before implementation. Please complete and submit a Request for Amendment Form available at the following website:

http://www.research.mq.edu.au/current_research_staff/human_research_ethics/managing_approved_research_projects

5. Please notify the Sub-Committee immediately in the event of any adverse effects on participants or of any unforeseen events that affect the continued ethical acceptability of the project.

6. At all times you are responsible for the ethical conduct of your research in accordance with the guidelines established by the University. This information is available at the following websites:

http://www.mq.edu.au/policy

http://www.research.mq.edu.au/for/researchers/how_to_obtain_ethics_approval/human_research_ethics/policy

If you will be applying for or have applied for internal or external funding for the above project it is your responsibility to provide the Macquarie University's Research Grants Management Assistant with a copy of this email as soon as possible. Internal and External funding agencies will not be informed that you have approval for your project and funds will not be released until the Research Grants Management Assistant has received a copy of this email.

If you need to provide a hard copy letter of approval to an external organisation as evidence that you have approval, please do not hesitate to contact the Ethics Secretariat at the address below.

Please retain a copy of this email as this is your official notification of ethics approval.

Yours sincerely,

Dr Anthony Miller
Chair
Faculty of Human Sciences
Human Research Ethics Sub-Committee
---
Faculty of Human Sciences - Ethics
Research Office
Level 3, Research HUB, Building C5C
Macquarie University
NSW 2109

Ph: +61 2 9850 4197
Email: fhs.ethics@mq.edu.au
http://www.research.mq.edu.au/