Cognitive load in simultaneous interpreting: Existing theories – New models

Article *in* Interpreting International Journal of Research and Practice in Interpreting · July 2011

DOI: 10.1075/intp.13.2.02see

Kilian G. Seeber
University of Geneva

John Benjamins Publishing Company

Cognitive load in simultaneous interpreting

Existing theories — new models

Kilian G. Seeber
University of Geneva

This paper sets out to describe and graphically illustrate the amount of cognitive load generated during the simultaneous interpretation of structurally different languages based on theories developed and evidence gathered in cognitive psychology and psycholinguistics. To that end, a German verb-final and verb-initial construction are analyzed and contrasted in terms of the load they cause to an inherently capacity-limited system when interpreted simultaneously into a verb-initial language like English. A series of analytical cognitive load models are introduced providing a detailed illustration of conjectured cognitive resource allocation during simultaneous interpreting of verb-final structures.

Keywords: cognitive load, working memory, capacity theory, attention, syntactic complexity

1. Introduction

In over half a century of research, simultaneous interpreting (SI) has been addressed by different authors from diverse disciplines with the goal of studying a cognitive task (Lederer 1978; Lonsdale 1997) that many identified as "complex" (Massaro & Shlesinger 1997; De Groot 2000), "difficult" (Gile 1995; Moser-Mercer 1997) and "demanding" (Hyönä et al. 1995; Rinne et al. 2000).

Even within the narrow field of interpreting research different schools of thought work within dissimilar paradigms and suggest quite diverse approaches to the study of one and the same object. Their approaches differ not only in the tools and methods they apply, but also in the degree to which they interact with other (arguably related) disciplines. Whereas the research community more akin to the natural sciences (Moser-Mercer 1994) appears to be rather permeable to influence from outside, willing to allow if not promote interdisciplinary work, the liberal arts-oriented community of researchers seems reluctant to embrace other paradigms.

Interpreting 13:2 (2011), **176–204**. DOI 10.1075/intp.13.2.02see
ISSN 1384–6647 / E-ISSN 1569–982X

Yet while caution might be called for when borrowing from neighboring disciplines, if it is true that, "much of what the simultaneous interpreter does is the same as we all do all the time" (Henderson in Shlesinger 2000: 1), it seems warranted to at least *start* exploring and describing SI using the tools and methods that have been tested and tried in the disciplines traditionally dealing with what *we all do all the time*, i.e., cognitive psychology and psycholinguistics. After all, in spite of its unique nature, from a procedural perspective, SI remains a language-processing task.

Against this background this contribution provides an overview of theories and empirical findings that have strongly influenced the academic discourse in disciplines adjacent to that of research in interpreting and that hold considerable potential to inform the latter. As not all of them apply directly or without reservations, an attempt is made to present competing accounts, and to elucidate some phenomena which remain unexplained. In particular, I will introduce a set of models in an attempt to illustrate the cognitive processes and workload involved during SI of selected language structures. In doing so I hope to revisit some of the arguments which have been brought forward but perhaps not sufficiently discussed in the body of literature addressing structural issues in SI.

The analysis begins with a discussion of the notion of working memory as a capacity limited system (Section 2). Section three provides a short summary of the involvement of working memory in language processing tasks whereas section four addresses issues pertaining to language comprehension. In section five we take a closer look at syntactic factors contributing to comprehension difficulties before introducing SI as a complex language processing task in section six. Section seven lays out the general theoretical foundations of the Cognitive Load Model of Simultaneous Interpreting and contrasts it with an existing model, i.e., Gile's (1995) Effort Model. It furthermore provides a detailed discussion of the model and of how it captures cognitive load for different interpreting strategies. The main features and predictions of the model are summarized in section eight.

2. Working memory

2.1 Working memory and attention

The idea of limited attentional resources goes back to "one of the most obvious behavioural properties of the human information processing system [which] is that there seems to be a fundamental limit on our ability to do a number of things at once" (Styles 1998: 116). In an attempt to substantiate a theory hitherto based mainly on introspection and observation, Welford (1952) presented participants with two stimuli in rapid succession, instructing them to respond immediately after stimulus presentation. When the second stimulus was presented prior to the subject's response to the first, the second response was delayed (Fagot & Pashler 1992). Welford interpreted the delay between the response to the first and second stimuli, also known as the PRP (Psychological Refractory Period), as corroborating evidence for a limited capacity mechanism that is able to process only one response decision at any given time. Based on these findings, Broadbent (1958) proposed his *filter theory* (using the "bottleneck" metaphor), suggesting the existence of one single channel of limited capacity that is able to handle a limited amount of information at any given time. Extensive experiments on dichotic listening revealed that participants' performance decreases when attention needs to be divided between two independent auditory input channels. Broadbent thus introduced the idea of the *filter* that is supposed to keep the system from overloading.[1] Due to its physical limitations, the filter cannot execute an infinite number of operations simultaneously. With the advent of computers, the image of the bottleneck gave way to this modern metaphor, although the latter may be equally simplistic and limited. In 1963, Knowles postulated the existence of a pool of limited capacity resources, presenting a theory according to which primary task workload correlates negatively with secondary task performance. Similarly, Kahneman (1973) put forward the existence of a central resource of capacity, which, like the processing capacity of a computer's CPU, can be deliberately allocated wherever necessary. The source of interference was thus no longer hypothesized as an impasse (i.e., a bottleneck), but as a particular stage of processing with high processing demands. Experimental research, however, did not support the hypothesis of a general pool of resources, as primary task workload *does not always* seem to affect secondary task performance (Allport et al. 1972; Wickens 1976; Schneider et al. 1984). The explanation for the inconsistency of the results of the dual-task experiment is two-fold: the degree of interference between or among simultaneously executed cognitive and motor tasks, which is believed to depend on the degree of automation (Schneider et al. 1984) of the individual tasks involved, on one hand, and on the structure of the tasks, on the other (Sanders 1979; Wickens 1984). Finally, Pashler and Johnston re-proposed the bottleneck theory, based on evidence suggesting that PRP effects persist in spite of structural dissimilarity between tasks, conceding that "it is quite possible that a number of bottlenecks will be discovered" (1998: 175).

It is important to keep in mind that the above claims pertaining to the nature of attention are made against the background of fundamental yet unresolved issues such as whether there are different attentional mechanisms underlying different tasks and modalities (Styles 1997), and whether attention is indeed shared when two tasks are carried out simultaneously (i.e., parallel processing), or whether it is switched between them (i.e., time-shared processing). Accordingly, Pashler and

Johnston's claim that "there is considerable evidence that a central bottleneck is a principle cause of PRP interference" (1998: 170) appears in contrast to more recent findings by Schumacher et al. (2001), who showed that "after relatively modest amounts of practice, at least some of the participants achieve virtually perfect time sharing in the dual-task performance of basic choice reaction tasks" (2001: 101).

2.2 Working memory and capacity

Working memory (WM) appears to be limited not only in the number of concurrent operations it can carry out, but arguably also in the amount of information it can keep available for processing. The quantification of this working memory capacity has been debated fervently for half a decade — ever since Miller's (1956) highly influential paper suggesting working memory span to be limited to a fixed number of so-called "chunks" of information. In spite of Miller's reluctance to provide a detailed definition of what actually constitutes a chunk (described rather generally as a set of items of information that are merged into one larger retrievable unit of information) the concept was adopted by various authors in the domain of both verbal (Miller 1956; Slak 1970; Simon 1974; Cowan et al. 2004) and non-verbal processing (De Groot 1965; Chase & Simon 1973; Gobet & Simon 1996a, 1996b, 1998) and compelling evidence was collected which supports the idea of a constant working memory capacity. Whereas Miller (1956) seemed to suggest a WM capacity limit of 7±2 chunks, other authors posited such capacity limit to be closer to three (Broadbent 1975) or four (Coltheart 1972; Cowan 2001). It should be noted that Miller himself (Miller 1989) has since argued that the number of chunks was used as a figure of speech.

Based on their study of expert chess players, Chase and Simon (1973) postulated their *chunking theory*, whereby chess masters and other experts acquire a large number of chunks both through practice and through study and the chunks can then be accessed through a discrimination net (on the basis of different perceptual features). This theory suggests general cognitive parameters, such as the time needed to learn a new chunk or the overall number that can be held in working memory, to be the same both for experts and for novices. The difference in performance is attributed to an expert's ability to recognize more and larger chunks, and thus find quicker, more appropriate responses. Drawing mainly on that theory, Gobet and Simon (1998) introduce their template theory, attributing expert performance to a large database of chunks indexed by a *discrimination net*, a large semantic memory with schemas and productions, and associations between the former and the latter. Chunks recurring regularly during practice or study are thought to evolve into more complex structures (templates), which allow information to be encoded into long-term memory (LTM) more swiftly. This approach is not entirely unlike Ericsson & Kintsch's (1995) Long Term Working Memory (LTWM) theory, with the difference that templates can only be accessed through perceptual cues. Cowan (2005) includes the notion of chunks[2] in his theory of *focused attention*, suggesting that chunking takes place at different levels and is organized hierarchically. Working memory may then shift between the different hierarchical levels in order to reduce load: chunks can be retrieved and merged into a larger chunk, thereby freeing up capacity (Cowan 2000, 2001). Although Cowan sees "no good reason to assume that the units of information in working memory are generally as simple and unconnected as chunks" (2005: 84), he believes Miller's notion of chunks to be useful as a discrete unit of measurement of a non-discrete entity such as the load on working memory.

As we shall see in the ensuing debate, the notion of chunks remains crucial in the discussion of language processing.

3. Language processing

Language processing refers to a range of sub-processes involving multiple levels of analysis (phonological, lexical, syntactic, semantic, and pragmatic) necessary to perform word recognition, parsing, information extraction, semantic interpretation, discourse analysis and ambiguity resolution — tasks which eventually enable us to extract meaning from what we hear, and to integrate it into a single coherent interpretation with incredible rapidity (Osterhout 1994). After a word has been recognized, a process shown to be very rapid — words spoken in context can be identified 200 ms after onset (Marslen-Wilson 1984) — effortless and devoid of conscious difficulty (Harley 2001), information about its meaning (i.e., semantic information) and about the rules governing its relationship to other words (i.e., syntactic information) becomes available. In order to establish the relationship between and among words, we assign thematic roles to them on the basis of a syntactic analysis (parsing). There has been an ongoing debate about the point in time at which non-structural (i.e., semantic, prosodic, contextual) information is integrated to construct a syntactic representation (see Harley 2001). Autonomous models of parsing such as Frazier's (1987) garden path model are largely based on Chomsky's claim that syntactic processes operate independently of others and are based on the principles of minimal attachment[3] and late closure,[4] and they identify two discrete stages of parsing (hence also *two-stage models*). Whereas the first stage draws exclusively on syntactic information, the second stage uses semantic and contextual information in order to build the syntactic representation of a sentence or clause. Although Fodor et al. (1974) suggest that only once a clause has been completely analyzed syntactically do we proceed to build a semantic representation of that clause, more recent proponents of these models argue that the second stage of parsing happens very quickly after the first. Conversely, interactive models of parsing (e.g. constraint-based models by Boland et al. 1990; Trueswell et al. 1993; Gibson 1998, 2000; Gibson & Pearlmutter 1998) assume that semantic information is immediately available to the syntactic processor and that four (or in the case of spoken language comprehension, five) main categories of constraints influence the parse: lexical and word-level constraints; contextual constraints (i.e., communicative utility, plausibility); computational resource constraints; and phrase-level contingent frequency constraints. Lately, interactive models have received considerable attention and a consensus seems to be forming in their favor (Harley 2001).

4. From language processing to comprehension

Comprehending a sentence involves a process of semantic and syntactic binding (Ruchkin et al. 2003). Once individual words of an utterance have been recognized, their thematic role identified and their meaning accessed, the listener integrates all this information into a representation, a mental model (see Johnson-Laird 1983; Garnham 1987) or a situational model (see van Dijk & Kintsch 1983) of the discourse. This model is a mental representation of the people, the objects, the locations and the events described in the discourse rather than its words, clauses, and sentences (Zwaan 1999). With every new bit of information, we add to and modify this representation, which according to Ericsson and Kintsch (1995) is stored in short term memory during the integration phase and in LTM once a sentence has been processed. Indeed, although Kintsch et al. (1999) clearly identify LTWM as an expert skill, they go on to claim that there are certain tasks, such as comprehension, "in which most adults in our society are experts" (1999: 4). Baddeley (2000), on the other hand, concludes that these structures are temporarily stored in the Episodic Buffer and are complemented by information retrieved in LTM. The process of comprehension, however, goes beyond the mere analysis of what we hear (or read). On the one hand, comprehenders recognize words, their thematic role and their meaning, and then attempt to establish relations between and among the individual parts of the text. On the other, they attempt to relate this information to their knowledge of the world in order to complement the mental model of the discourse. Both are instantiations of a process known as inferencing.

The comprehension process of narrative discourse requires comprehenders to mentally link successive events so as to form a coherent representation (mental model) of the story. During this process, comprehenders use the information provided by speakers in order to convey their intentions in order to infer the latter

(Wilson 1999). Comprehension is thus based on the interaction between information that is explicitly stated in the form of (written) text or (spoken) discourse, and the information comprehenders bring to bear (Garrod et al. 1990). Although various authors (e.g. Clark 1975; Harris & Monaco 1978; Garrod et al. 1990; Van den Broek 1994) have described and categorized several different kinds of inferences, there is little agreement on the exact nature of the particular information to which comprehenders gain access during the comprehension process. It seems to be generally accepted, however, that procedural and computational constraints and resources are a limiting factor to inference processing (Van den Broek 1994). Whereas the spread of activation is claimed to be automatic (Heil et al. 2004), and thus not limited by capacity constraints (Schneider et al. 1984), all search and match processes during inferencing (whether intentional or automatic) are posited to be resource-consuming (Wilson 1999). Consequently, the comprehension process (e.g. during reading) will only be swift and smooth so long as the inferential processes required for the construction of a coherent representation do not exceed the available memory resources. In summing up, it seems fair to say that the inferencing process is influenced by a number of factors, above all by the comprehenders themselves, as they vary in their working memory capacity, their verbal ability, the quality and quantity of their knowledge and their comprehension goals (Singer 1994). The comprehenders' background knowledge influences the inferencing process insofar as it guides construction of the macrostructure (see Kintsch & van Dijk 1978) and sometimes the "comprehenders' preferences for a particular outcome of a story interfere with the verification of previously known information about the actual outcome of the story" (Zwaan 1999: 17). Similarly, task demands and inference processing are believed to interact closely (Singer 1994), meaning that demands imposed by the task may reduce the comprehender's processing capacity, thus limiting inference processing (e.g. when time constraints do not allow the comprehender to execute all the processes required for comprehension). Conversely, the comprehender may not engage in inferential processing when the task does not demand such processing. In fact, some processing tasks such as problem solving or learning may require deeper processing (Craik & Lockhart 1972), or simply entail the construction of different kinds of internal representations (Mayer 1983) than others, such as summarizing or listening for pleasure.

5. Syntax and processing difficulty

Although it has been suggested that the integration of incoming linguistic input into an existing structure is effortless most of the time (Ferreira et al. 2002; Kamide et al. 2003), in a limited-resource paradigm this process is hypothesized to require a certain (even if unconscious) cognitive effort. Friederici and Bornkessel believe that "syntactic aspects of working memory are of major relevance during sentence comprehension, in particular when the sentence is syntactically complex" (2003: 763). Already in the sixties Yngve (1960) and Chomsky and Miller (1963) identified particular syntactic phenomena, such as nesting or center-embedding, that were more difficult to process and understand (as reflected by longer reading times) than left- or right-branching structures, the rationale being that the former tax syntactic storage more than the latter. However, there is still little accord on which aspects of syntactic structure account for storage cost and how such storage cost can be quantified. Whereas Yngve (1960) and Chomsky and Miller (1963) propose quantifying syntactic storage at any given point of a parse in terms of the number or partially processed phrase structure rules, Kimball (1973) suggests incomplete clauses as units of syntactic storage, Stabler (1994) argues in favor of case assignments, and Gibson (1998, 2000) makes his case for predicted syntactic heads. Unlike other theories (such as minimal attachment), which account for increased reading or comprehension times (i.e., increased processing requirements) in ambiguous structures but fall short of explaining the phenomenon in unambiguous structures, Gibson's dependency locality theory (DLT) explains both accounts. According to his theory, there are two important components of sentence parsing that consume computational resources. First, whenever a new word has to be integrated into an existing structure, this will entail structural integration cost. Second, whenever incomplete dependencies need to be kept track of, this will entail memory cost. As such, the DLT is closely related to the claims of working memory as a system with both storage and processing components (cf. Baddeley 2000). Structural integration cost is assumed to depend on the locality (i.e., distance) between the two elements that need to be integrated, and although distance can be expressed in different ways — Gibson (1998) measures distance in terms of complexity of the intervening discourse structure, Hawkins (1994) in terms of the number of words, Lewis (1996) and Gordon et al. (2001) in terms of the number of interfering similar elements — the predictions of the DLT are the same regardless of which measure of distance is applied (see Grodner & Gibson 2003). Following this rationale, Gibson argues that in English, long-distance connections are more difficult to integrate than shorter ones. In head-final structures like those used in Japanese or German, however, very few, if any, effects of integration cost were found at the end of such sentences. In fact, it appears as though even sentences whose main verb, i.e., the most important semantic and syntactic constituent, is dislocated, are easy to comprehend (Scheepers et al. 1999). Such findings challenge locality-based predictions and favor an interpretation according to which subsequent items are anticipated through the integration and projection of previous items. In other words, if the verb is placed at the end of a sentence, it might be anticipated through the number and type of arguments preceding it (see Konieczny 1996; Konieczny & Hemforth 1994). Furthermore, evidence suggests that pre-head integration of arguments takes place incrementally, meaning that in order to be integrated into the structure, previous, unattached items do not need to be attached to the head, i.e., the verb, (Hemforth et al. 1993; Bader & Lasser 1994; Kamide & Mitchell 1999).

5.1 Processing difficulty of verb-final structures

Let us take a closer look at German, which features a very flexible word order, including the position of the verb. In German subordinate clauses all main verbs appear at the end of the clause. In declarative main clauses, however, simple tense verbs appear in second position (V2), whereas compound tense verbs are separated: the finite verb component (e.g. an auxiliary) appears in V2 and the non-finite verb component (e.g. a participle) appears at the end of the clause (VF). Beyond that, there is a V2 effect whenever a constituent other than the subject is moved to the front of the clause, i.e., the finite verb component moves into second position preceding the subject (now in third position). The absence of any significant integration cost at the end of these constructions may be accounted for by the fact that verb-final structures in German fulfill the expectations with regard to the order of arguments and can therefore be integrated directly into the phrase structure (see Fiebach et al. 2001). Also, although one could expect storage effects to be considerable, the processing of these sentences is often no slower towards the end; rather, it may be faster, conceivably because other factors narrow the choice of which verb to expect. Konieczny and Döring (2003) draw on MacDonald and Christiansen's (2002) Simple Recurrent Network (SRN) to account for subjects' ability to anticipate German clause-final heads. In other words, the more exposure the network receives to particular linguistic regularities, the more it is trained to predict upcoming constituents. As German verb-final constructions exhibit regular word order, it is assumed that the more preceding dependents the sentence features, the easier the processing will be. Although Gibson's (1998) DLT does not account for Konieczny and Döring's results, it is important to point out that Gibson's (1998, 2000) and Grodner and Gibson's (2003) experiments were carried out using a single sentence paradigm, thereby perhaps limiting the influence of some of the other factors identified by Gibson and Pearlmutter as "interacting informational constraints" (1998: 262). Processing discourse or a sequence of sentences is possible in spite of the limitations of working memory because storage demands are minimized through immediate processing (i.e., interpreting each new word or phrase as far as possible upon first encounter) and because context provides processing benefits (Just & Carpenter 1992). What is more, Gibson's experiments used English stimuli, which by definition lack some of the compensatory features used by many languages with SOV structure (e.g. Japanese or German) to offset the conjectured load imposed by the dislocated main verb, such as heavy case markings and inflections (Ueno & Polinsky 2005). This means then that whereas a simplistic version of Gibson's DLT predicts more complexity for the sentence (A) *Ich glaube, dass die Delegierten ihre Entscheidung nach einer langen Debatte treffen* than for the sentence (B) *Ich glaube, die Delegierten treffen ihre Entscheidung nach einer langen Debatte*, Scheepers et al. (1999), Konieczny (1996) and Konieczny and Hemforth (1994) do not predict a difference in processing load attributable to the locality of the verb. Provided there are sufficient informational constraints, they would argue that the integration and memory cost at the verb *treffen* is no larger in sentence A than in sentence B.

6. SI as a complex language processing task

Simultaneous interpreting is the process of cross-linguistic transfer of meaning in real time. From an information-processing perspective the notion of *real time* deserves particular attention. Indeed the term "simultaneous" in SI does not imply the simultaneity of the comprehension and production of one and the same sentence constituent, but the general temporal overlap of language comprehension and language production (Christoffels 2004). In fact, given that the interpreter generally needs to listen to the SL discourse and comprehend it before venturing to encode and produce the TL discourse (with the exception of instances in which the interpreter overtly anticipates the speaker), the transfer as such can hardly be simultaneous. The average ear-to-voice span (EVS), i.e., the time lag between the speaker and the interpreter, was found to be around 3 seconds (albeit with a considerable range of over 10 seconds) in early accounts of SI (Oléron & Nanpon 1965; Barik 1973) as well as in more recent studies (Lee 2002; Christoffels 2004). However, variables such as the language combinations involved as well as directionality and the nature of the materials may influence the EVS. These variables may account for the differences in average EVS which Oléron and Nanpon found between interpreters working from German into French (1.9 seconds), from English into French (2.6 seconds) and from French into English (5.4 seconds). Similarly, they may explain Goldman-Eisler's observation that "when translating from German, interpreters delay translation longer than when translating from French or English," which is in stark contrast to Oléron and Nanpon's results, and which she believes is "most probably because the predicate in German comes at the end of the proposition." (1972: 136). It should also be pointed out that whereas Lee's (2002) results reflect the performance of professional interpreters working from

L2 English into L1 Korean, Christoffels' (2004) data were gathered using untrained bilinguals, working from Dutch L1 into English L2 and vice versa.

6.1 Absolute and relative problems in language processing

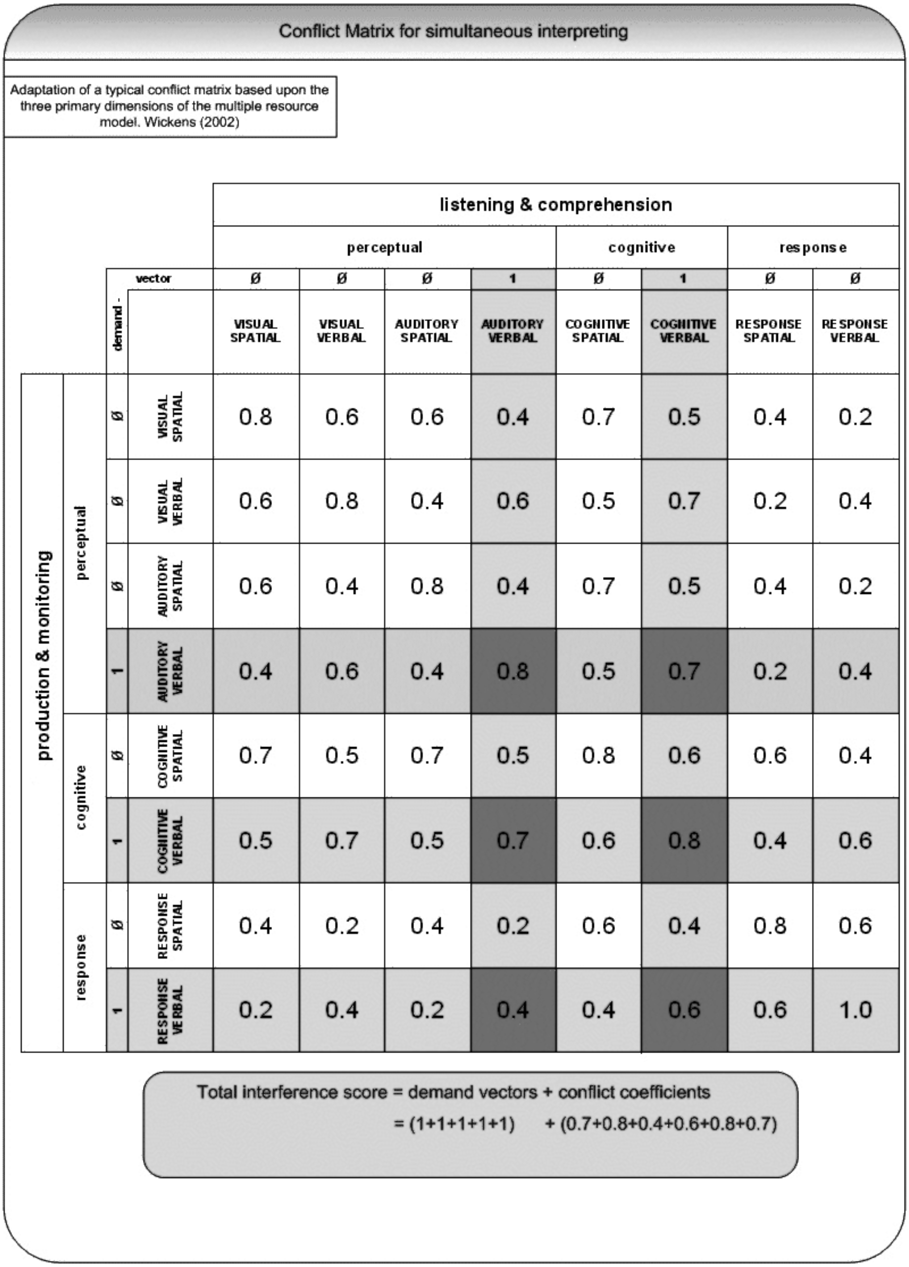
It appears warranted to assume that certain factors limiting language comprehension and production will persist when these tasks are combined into an arguably more complex task such as SI. On the other hand, the simultaneous performance of several cognitive tasks is likely to reveal new constraints which, rather than being inherent in the component tasks, do not emerge until they are combined, or else, their effect is negligible within one task but is compounded when processes are combined. This view is at odds with authors who believe that "only factors which impair normal comprehension should impair SI" (Setton 1999: 54). Among the prime examples of such limiting factors is the input rate of the source discourse, which has been studied in more detail than factors such as semantic density, lexical suppression, self-monitoring, and syntactic complexity, the effects of which on the simultaneous interpreting task have yet to be substantiated empirically. Although experts seem to agree that the normal public speaking rate ranges between 100 and 200 wpm (Mayer 1988; Verderber 1984; DeVito 1981), experimental evidence suggests that "when all other variables are controlled, the rate of delivery, whether hesitant (90 to 100 wpm) or rapid-fire (350 to 500 wpm, rates frequently attainable only by mechanical manipulation of pre-recorded material), does not significantly affect comprehension" (Voor & Miller 1965: 452). In spite of these findings, the input rate recommended for SI by advocates of different schools of thought ranges between 95 and 120 wpm (Gerver 1976; Seleskovitch 1978; Lederer 1981). These recommendations seem warranted: research indicates that omissions, substitutions and pronunciation errors (Pio 2003) increase with higher rates of input,[5] whereas anticipation accuracy decreases (Seeber 2005). Such findings lend strong support to the view that whereas input rate may not constitute an absolute problem for comprehension (at least not before age-related problems of comprehension appear), it becomes a major constraining factor during SI.

The importance of language-specific factors (i.e., the fact that some language pairs are structurally very different, resulting in varying degrees of syntactic asymmetry between the SL and the TL) for the SI process remains a hotly debated issue, dividing the interpreting research community. There are those who would claim that the SI process is unaffected by the canonical word order of the two languages, based on the fact that native speakers of verb-final languages like German do not seem to have any problem understanding such (SOV) structures and often even seem to be able to anticipate the verb (cf. Lederer 1981; Seleskovitch 1984; but also Konieczny & Hemforth 1994; Konieczny 1996). However, this claim seems tenuous, particularly since it has yet to be substantiated in an SI paradigm and does not consider the temporal dynamics of the modern interpreter's working environment. Even so, professional conference interpreters seem to have acquired strategies for dealing with temporal constraints, as I will try to demonstrate by means of the Cognitive Load Model.

7. Towards a Cognitive Load Model of SI

As we have seen in the preceding sections, simultaneous interpreting is an instantiation of multitasking that requires the interpreter to engage in a language comprehension task and a language production task at the same time. In cognitive processing terms, the real-time combination of the two means that they compete for available resources. Given that single-resource theories (e.g. Kahneman 1973) do not account for perfect time-sharing between tasks and cannot explain why a change in task structure (at constant difficulty) results in different degrees of interference between tasks, the model to be developed here is based on Wickens' (1984) Multiple Resource Model, in which the combination of two (or more) tasks requires more processing capacity than either (or any) of the tasks performed individually. More importantly, and unlike Kahneman's (1973) single resource theory, which assumes that all cognitive tasks compete for one undifferentiated pool of resources, the Multiple Resource Model assumes that tasks interfere with each other more strongly when they have structures in common, i.e., if they demand the same level of a particular processing dimension, than if they rely on different structures. The model therefore predicts substantial interference between resource-demanding perceptual tasks and cognitive tasks involving working memory to store or transform information (Liu & Wickens 1992; Wickens 2002), as is the case in language production and perception. The conflict matrix (see Figure 1), a computational model of multiple resources predicting the amount of interference between different cognitive tasks, complements the Multiple Resource Model. This matrix will be used to analyze the amount of interference generated by the simultaneous performance of the aforementioned tasks.

In order to capture the notion of cognitive load in SI, I propose a Cognitive Load Model that takes into account the amount of load generated by individual concurrent tasks. For that purpose, SI will be considered a real-time combination of a language comprehension and a language production task (see discussion above). Both tasks are broken down into their demand vectors (i.e., perceptual auditory verbal processing of input and output 'P', cognitive-verbal processing of input and output 'C', and verbal-response processing of output 'R') and interference 'I' is calculated (and added as a conflict coefficient) whenever two or more



Conflict Matrix for simultaneous interpreting

Adaptation of a typical conflict matrix based upon the three primary dimensions of the multiple resource model. Wickens (2002)

				listening & comprehension							
				perceptual				cognitive		response	
			vector	Ø	Ø	Ø	1	Ø	1	Ø	Ø
		demand -		VISUAL SPATIAL	VISUAL VERBAL	AUDITORY SPATIAL	AUDITORY VERBAL	COGNITIVE SPATIAL	COGNITIVE VERBAL	RESPONSE SPATIAL	RESPONSE VERBAL
production & monitoring	perceptual	Ø	VISUAL SPATIAL	0.8	0.6	0.6	0.4	0.7	0.5	0.4	0.2
		Ø	VISUAL VERBAL	0.6	0.8	0.4	0.6	0.5	0.7	0.2	0.4
		Ø	AUDITORY SPATIAL	0.6	0.4	0.8	0.4	0.7	0.5	0.4	0.2
		1	AUDITORY VERBAL	0.4	0.6	0.4	0.8	0.5	0.7	0.2	0.4
	cognitive	Ø	COGNITIVE SPATIAL	0.7	0.5	0.7	0.5	0.8	0.6	0.6	0.4
		1	COGNITIVE VERBAL	0.5	0.7	0.5	0.7	0.6	0.8	0.4	0.6
	response	Ø	RESPONSE SPATIAL	0.4	0.2	0.4	0.2	0.6	0.4	0.8	0.6
		1	RESPONSE VERBAL	0.2	0.4	0.2	0.4	0.4	0.6	0.6	1.0

Total interference score = demand vectors + conflict coefficients
= (1+1+1+1+1) + (0.7+0.8+0.4+0.6+0.8+0.7)

Figure 1. Conflict matrix for simultaneous interpreting

tasks overlap. We furthermore include a storage component 'S' that reflects the load generated by the storage in working memory of constituents prior to their integration and/or production. This load component is added to that of language comprehension because the nature of SI is likely to entail greater memory demands than ordinary language comprehension (Nelson Cowan, personal communication).

7.1 The Cognitive Load Model vs. the Effort Model

The Cognitive Load Model was inspired by Gile's (1995) Effort Model of simultaneous interpreting. The two models share the same objective to the extent that both attempt to capture and illustrate the cognitive demands inherent to SI. The theoretical foundation of the Cognitive Load Model, however, is in stark contrast to that of the Effort Model. The latter is based on Kahneman's (1973) single resource theory and thus assumes that all tasks involved in the SI process draw on one and the same pool of undifferentiated resources. Not only does this approach preclude the model from accounting for well-documented phenomena such as perfect time-sharing (Schumacher et al. 2001), it also implicitly assumes that resources can be shifted and re-allocated between or among tasks, which finds little support in the literature.

While both models conceive of SI as a combination of a language comprehension task and a language production task, suggesting an increment in overall cognitive load (or invested effort, as the case may be), only the Cognitive Load Model is able to account for the conflict potential posed by an overlap and the interference they cause.

Among the strengths of the Cognitive Load Model is its ability to reflect local cognitive load as a function of both input *and* output features. Demand vectors of both the original input and the interpreter's output are accounted for and provide a more detailed analysis of local cognitive load than the Effort Model. The graphic representation of the latter remains underspecified and provides only a tenuous link between the alleged comprehension and production efforts (see Gile 1995, 1997). What is more, unlike the Effort Model, the Cognitive Load Model also includes a first attempt at quantifying cognitive load, relying principally on Wickens' demand vectors and conflict coefficients.

Finally, Gile only very briefly addresses the issue of syntactic asymmetry in SI, indicating that "differences between the syntactic structures of the SL and the TL can increase the memory effort's processing capacity requirements because of the waiting involved before being able to reformulate the SL segment into the TL" (1997: 206). The Cognitive Load Model, on the other hand, illustrates how the overall cognitive demands are affected by the different combinations of sub-tasks, reflecting different strategies for coping with syntactic asymmetry between SL and TL. The Cognitive Load Model thus presents a more comprehensive view, including a strategic component and its repercussions on local and overall changes in cognitive load. Given these differences, I conceive of the Cognitive Load Model of SI as a competing account to Gile's Effort Model.

7.2 Cognitive load management at the macro level

I have argued that simultaneous interpreting combines two natural, robust language processing tasks, language comprehension and language production, into a more complex task. The combination of these tasks means that unlike during natural language production, production in SI starts from the intent to express an idea formed not by the speaker, but by someone else. Another difference lies in the fact that interpreters might have to react to less than a complete sentence or clause. It is plausible that when combined in real time, the processes underlying the two tasks allow synergies between them to be exploited and shortcuts to be taken in order to minimize the overall cognitive load inherent in both. This view is supported in the literature (Paradis 1994; Fabbro & Gran 1994; Isham 1994; Massaro & Shlesinger 1997), which reports two distinct yet complementary (Christoffels 2004) strategies used by the interpreter: meaning-based strategies and transcoding. Whereas the former implies that every part of the input is mediated through the conceptual stage (identified as the end point of the comprehension process and the start of the production process, respectively; see Moser 1978), the latter assumes the possibility of shortcuts at the discrete levels of language processing, based on the rationale that the interpreter's level of processing will be no deeper than needed (see Riccardi 1998; Alexieva 1998). Dillinger (1994) shows a joint effect of text type (narrative or procedural) and text structure (e.g. clause and propositional density, embedding, directness of mapping), suggesting that professional interpreters may opt for one or the other macro-strategy depending on the organization of propositional information in the source text.

7.3 Cognitive load management at the micro level

Against the background of this principle of economical processing, it appears reasonable to assume that the output produced by simultaneous interpreters is not haphazard but reflects particular behavioral patterns that have been acquired and are applied in an attempt to deal with the constraints inherent to the task (Riccardi 1998) and to save processing capacity (Dawrant 1996). It is furthermore conceivable that when given the choice, the interpreter will opt for a strategy that reduces overall cognitive processing demands. Whereas Gile (1999) suggested that simultaneous interpreters work close to cognitive saturation most of the time, local cognitive resource demands may in fact vary considerably, and professionally trained interpreters may well work with enough slack to allow them to overcome local impasses. In terms of syntax, this would mean that "when the word order of the TL parallels that of the SL, the interpreter is assumed to prefer a more or less left-to-right sequence; i.e., to use the least demanding strategy for the task" (Shlesinger 2000: 4). If we take this rationale a step further, one could argue that a Cognitive Load Model of SI of an SVO structure into another SVO structure (or any other syntactically symmetrical pair of structures) should provide us with a baseline in terms of cognitive requirements attributable to syntax for the simultaneous interpretation into English of German sentence B in the examples above. The model is based on the assumption that the constituents of an SL proposition have to be stored in a limited-resource WM system until they can be integrated into a situational model and eventually encoded in the TL. Given that this analysis focuses on the amount of cognitive load generated by structural features of the input and the output, the load generated by the comprehension and production of different sentence constituents

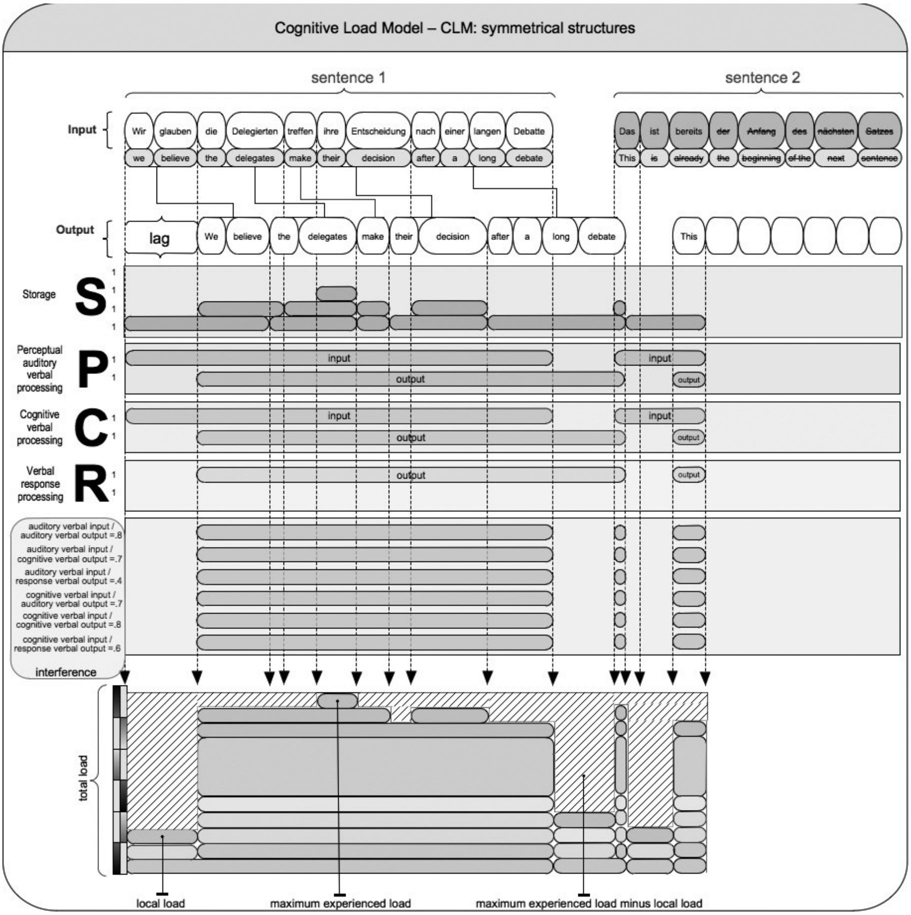

Cognitive Load Model – CLM: symmetrical structures
sentence 1
sentence 2
Input
Wir glauben die Delegierten treffen ihre Entscheidung nach einer langen Debatte
we believe the delegates make their decision after a long debate
Das ist bereits ~~der Anfang des nächsten Satzes~~
This ~~is already the beginning of the next sentence~~
Output
lag
We believe the delegates make their decision after a long debate
This
Storage
S
Perceptual auditory verbal processing
P
input
output
Cognitive verbal processing
C
input
output
Verbal response processing
R
output
auditory verbal input / auditory verbal output =.8
auditory verbal input / cognitive verbal output =.7
auditory verbal input / response verbal output =.4
cognitive verbal input / auditory verbal output =.7
cognitive verbal input / cognitive verbal output =.8
cognitive verbal input / response verbal output =.6
interference
total load
local load
maximum experienced load
maximum experienced load minus local load

Figure 2. Cognitive Load Model for symmetrical structures

is regarded as a constant. In other words, the present illustration does not account for factors influencing retrieval and activation mechanisms (e.g. semantic priming, relative frequency, abstractness, cognates and false cognate, etc.).

With these limitations in mind the Cognitive Load Model can illustrate the amount of cognitive load generated during the different interpreting solutions.

The CLM in Figure 2 illustrates the cognitive load experienced during SI of a German structure that allows interpreters working into English to maintain the syntax of the SL. The first-in-first-out processing of the sentence shows that the overall cognitive demands remain fairly constant once production begins and slowly decrease after the completion of the comprehension task for that particular proposition. Crucially, interpreters are able to finish producing their interpretation with a relatively short lag, allowing them to listen to the beginning of the next proposition without covering it up with their own production (see Gile 2008

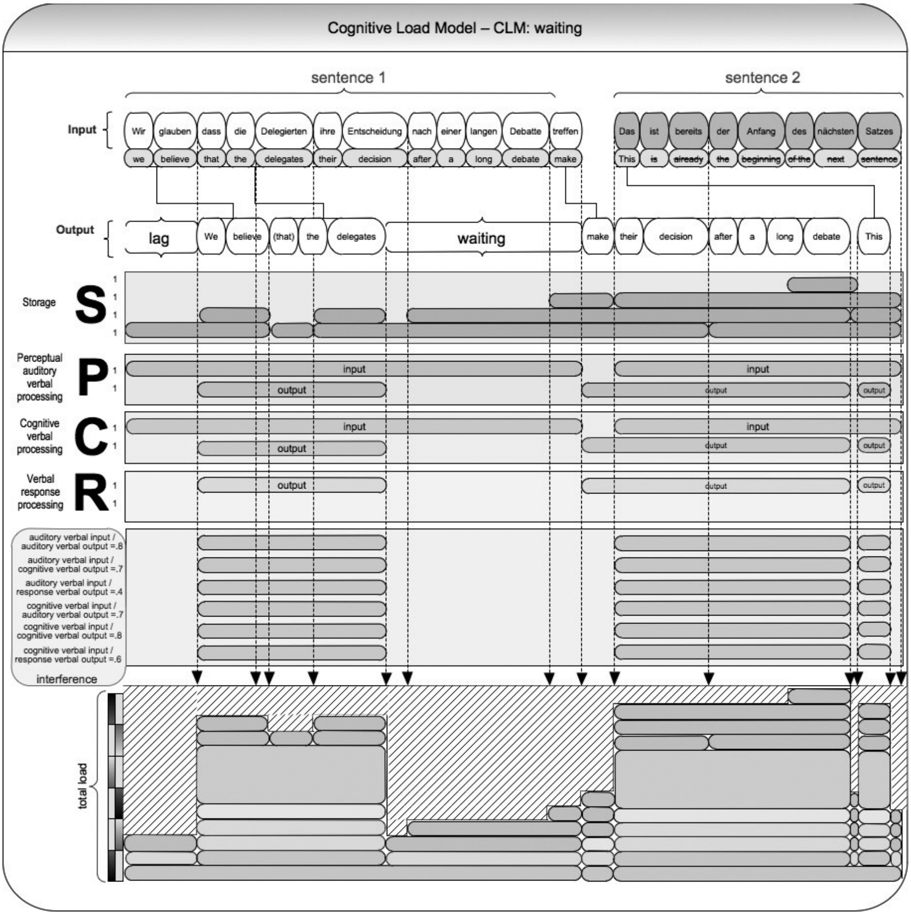

Cognitive Load Model – CLM: waiting
sentence 1
sentence 2
Input
Wir glauben dass die Delegierten ihre Entscheidung nach einer langen Debatte treffen
we believe that the delegates their decision after a long debate make
Das ist bereits der Anfang des nächsten Satzes
This ~~is~~ ~~already~~ ~~the~~ ~~beginning~~ ~~of the~~ ~~next~~ ~~sentence~~
Output
lag
We believe (that) the delegates
waiting
make their decision after a long debate
This
Storage
S
Perceptual auditory verbal processing
P
input
output
Cognitive verbal processing
C
Verbal response processing
R
auditory verbal input / auditory verbal output =.8
auditory verbal input / cognitive verbal output =.7
auditory verbal input / response verbal output =.4
cognitive verbal input / auditory verbal output =.7
cognitive verbal input / cognitive verbal output =.8
cognitive verbal input / response verbal output =.6
interference
total load

Figure 3. Cognitive Load Model — waiting

on *exported load*), which is likely to increase interference between the two tasks (Wickens 1984, 2002).

When working between syntactically asymmetrical language pairs such as German and English, however, the first-in-first-out strategy will not always work (e.g. in the case of German subordinate constructions). In fact, "differences in canonical word order […] and norms which favor an output that is 'acceptable' […] tend to rule out such a strategy, insofar as it is likely to produce a deviant word order in the TL" (Shlesinger 2000: 40). There are several ways in which simultaneous interpreters can deal with this constraint, four of which will be illustrated using sentence A from the example above and discussed briefly in terms of the cognitive load they produce: waiting, stalling, chunking and anticipating.

Waiting is the strategy by which simultaneous interpreters halt TL production to wait for more SL input. As a consequence, the information they receive whilst waiting must be stored in WM, where it needs to remain activated through rehearsal (a process which requires cognitive resources and is highly sensitive to interference from concurrent language processing tasks; see Wickens 2002) until it can be encoded in the TL. On one hand, this allows interpreters to alleviate cognitive load temporarily, as the interruption of simultaneous language comprehension and production effectively transforms the process into a simple comprehension and memorization task. On the other hand, however, it may cause a spillover effect, leading to a considerable increase in cognitive load downstream, i.e., when the information the interpreter was waiting for is finally encoded. The CLM indicates this increase in load and shows that by applying this strategy the interpreters accumulate a substantial EVS (Figure 3), thus delaying the load until the arrival of the next ST proposition.

As a strategy, *stalling* is very similar to waiting, as both aim to buy time, during which the interpreter may receive more input before the integration and encoding stage. Whereas waiting normally results in a period of silence in the interpreters' output, which may be perceived as uncomfortable by the listener and/or the interpreter, stalling postulates the production of "neutral padding" (see Figure 4) which fills the gap without adding any new information, although strictly speaking even the repetition of what has already been said alters the message slightly in terms of emphasis, etc.. Similar to the waiting strategy, stalling increases the interpreter's lag and adds a layer of processing complexity as the encoding and production of the padding material overlaps with the comprehension process. The CLM for the stalling strategy clearly indicates both the marked increase in cognitive load during padding as well as the overall accumulation of lag.

Text *chunking*, the comprehension process whereby a sentence is divided into smaller segments based on a superficial analysis, has been suggested by Abney (1991) as a precursor to full parsing. Chunks are posited to correspond roughly

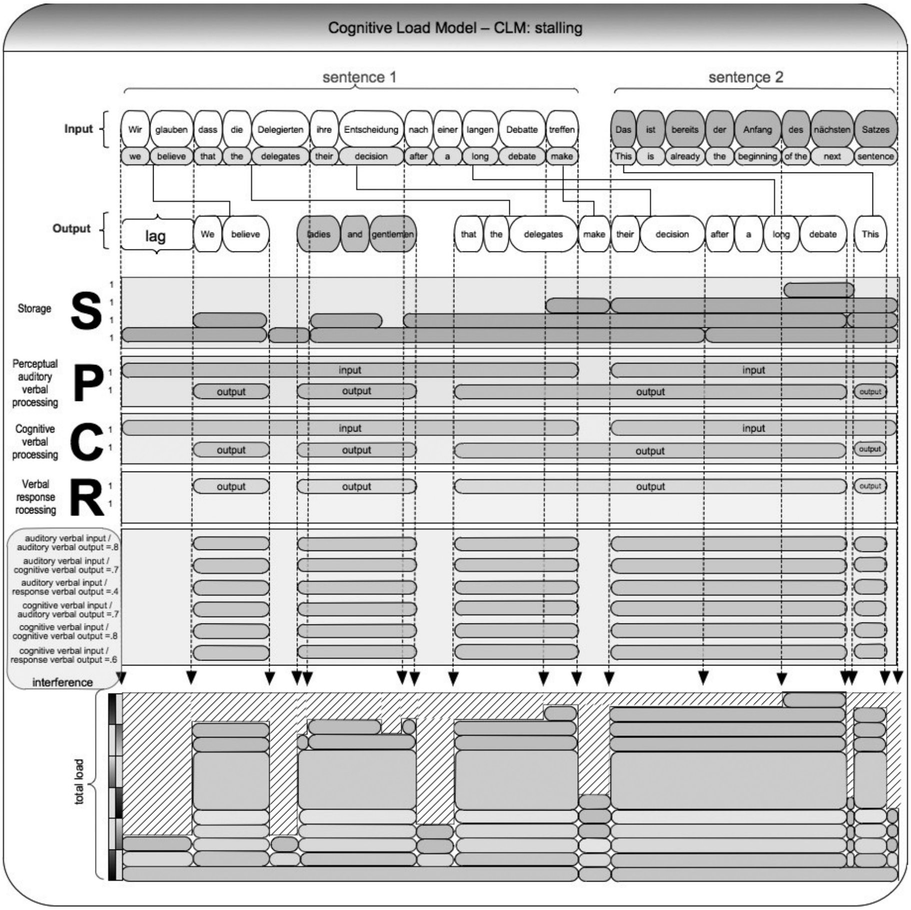

Cognitive Load Model – CLM: stalling
sentence 1
sentence 2
Input
Wir glauben dass die Delegierten ihre Entscheidung nach einer langen Debatte treffen
we believe that the delegates their decision after a long debate make
Das ist bereits der Anfang des nächsten Satzes
This is already the beginning of the next sentence
Output
lag
We believe
ladies and gentlemen
that the delegates make their decision after a long debate
This
Storage S
Perceptual auditory verbal processing P
Cognitive verbal processing C
Verbal response rocessing R
input
output
auditory verbal input / auditory verbal output =.8
auditory verbal input / cognitive verbal output =.7
auditory verbal input / response verbal output =.4
cognitive verbal input / auditory verbal output =.7
cognitive verbal input / cognitive verbal output =.8
cognitive verbal input / response verbal output =.6
interference
total load

Figure 4. Cognitive Load Model — stalling

to the prosodic patterns of a language and usually consist of a single content word surrounded by a number of function words. The advantage of chunking is that it allows for immediate integration of arguments and the postponing of more complex attachment decisions (Ramshaw & Marcus 1995). Similarly, as a strategy in SI, chunking refers to the process whereby interpreters segment the input into smaller fragments that can be encoded without having to wait for the entire sentence to unfold (see Figure 5). Although the SL input can be integrated and encoded immediately, the absence of a main verb relating the arguments to each other means that the chunks often need to be strung together downstream in order to establish (or recover) the original meaning, causing a temporally deferred increase in cognitive load. A non-negligible drawback of this strategy is that the resulting constructions are sometimes convoluted, or "extrêmement lourdes, [au point] qu'elles font carrément violence à la langue" (Ilg 1959: 10).

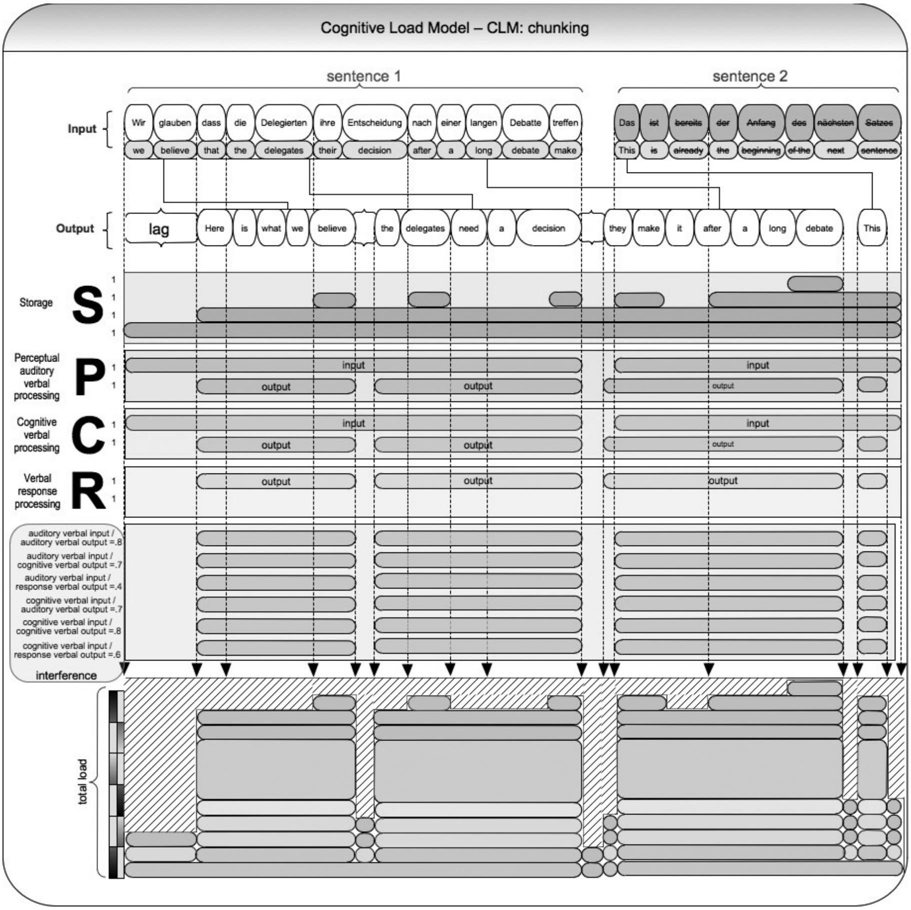

Cognitive Load Model – CLM: chunking
sentence 1
sentence 2
Input
Wir glauben dass die Delegierten ihre Entscheidung nach einer langen Debatte treffen
we believe that the delegates their decision after a long debate make
Das ~~ist~~ ~~bereits~~ ~~der~~ ~~Anfang~~ ~~des~~ ~~nächsten~~ ~~Satzes~~
This ~~is~~ ~~already~~ ~~the~~ ~~beginning~~ ~~of the~~ ~~next~~ ~~sentence~~
Output
lag
Here is what we believe
the delegates need a decision
they make it after a long debate
This
Storage
S
Perceptual auditory verbal processing
P
Cognitive verbal processing
C
Verbal response processing
R
1
input
output
auditory verbal input / auditory verbal output =.8
auditory verbal input / cognitive verbal output =.7
auditory verbal input / response verbal output =.4
cognitive verbal input / auditory verbal output =.7
cognitive verbal input / cognitive verbal output =.8
cognitive verbal input / response verbal output =.6
interference
total load

Figure 5. Cognitive Load Model — chunking

The fourth strategy, i.e., *anticipation*, refers to the interpreter's ability to predict a part of the original discourse before it has been uttered by the speaker, and has received considerable attention in the literature (Moser 1976; Kirchhoff 1976; Ilg 1959, 1978; Wilss 1978; Lederer 1981; Seleskovitch 1984; Van Dam 1989; Chernov 1992; Gile 1992; Kohn & Kalina 1996; Riccardi & Snelling 1997; Massaro & Shlesinger 1997; Zanetti 1999; Setton 1999; Seeber 2001, 2005). The CLM for this strategy (see Figure 6) suggests two significant advantages over the remaining three strategies. On the one hand, cognitive resource demands appear to remain close to baseline values with the exception of the actual inference processing (i.e., the 'guessing' of the verb), which is believed by many to recruit cognitive resources (see discussion above). On the other hand, this strategy allows interpreters to finish their interpretation with a lag similar to the baseline value, thus avoiding the aforementioned spillover effect. These two advantages make anticipation an

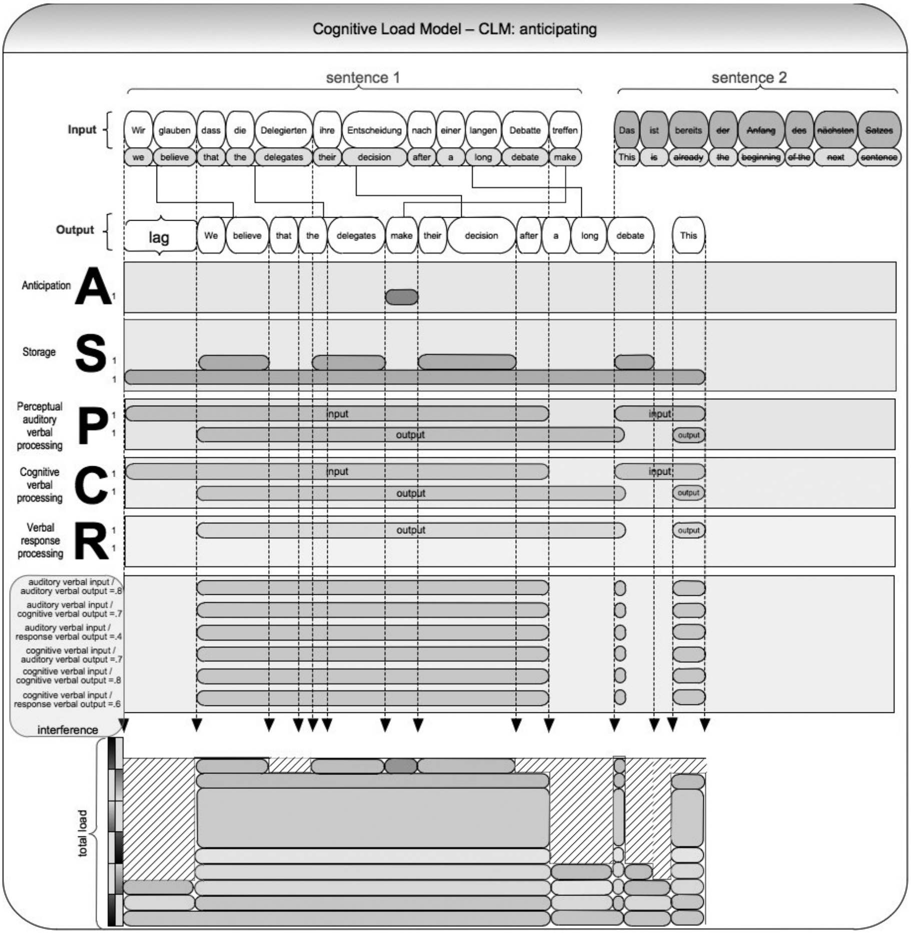

Cognitive Load Model – CLM: anticipating
sentence 1
sentence 2
Input
Wir glauben dass die Delegierten ihre Entscheidung nach einer langen Debatte treffen
we believe that the delegates their decision after a long debate make
Das ist bereits ~~der~~ ~~Anfang~~ ~~des~~ ~~nächsten~~ ~~Satzes~~
This ~~is~~ ~~already~~ ~~the~~ ~~beginning~~ ~~of the~~ ~~next~~ ~~sentence~~
Output
lag
We believe that the delegates make their decision after a long debate
This
Anticipation A 1
Storage S 1 1
Perceptual auditory verbal processing P 1 1
input
output
Cognitive verbal processing C 1 1
input
output
Verbal response processing R 1 1
output
auditory verbal input / auditory verbal output =.8
auditory verbal input / cognitive verbal output =.7
auditory verbal input / response verbal output =.4
cognitive verbal input / auditory verbal output =.7
cognitive verbal input / cognitive verbal output =.8
cognitive verbal input / response verbal output =.6
interference
total load

Figure 6. Cognitive Load Model — anticipating

ideal solution when interpreting German VF structures into English, yet empirical findings (Seeber 2001) corroborate Ilg's warning about its inherent danger: "cette technique est extrêmement dangereuse, car souvent quelque verbe-surprise vient déjouer les prévisions. L'effet stylistique et psychologique de ces verbes est très grand, puisque tout le poids de la phrase s'y trouve concentré. Il ne faut donc pas le dénaturer" (1959: 9).

8. Conclusion

The Cognitive Load Models for the different SI strategies provide a detailed illustration of conjectured local cognitive load. Given the analytical nature of the model, however, I will comment only on the three principal trends that seem to emerge from this analysis.

First, as compared to the baseline (i.e., the symmetrical condition), all interpreting strategies, with the exception of verb anticipation, entail a considerable increase in lag. This 'spillover effect' has already been addressed by Gile (2008), who believes that it might be the explanation for an increase in cognitive load downstream which may lead, in turn, to process breakdown. The present analysis corroborates his claim.

Second, cognitive load appears to vary according to the micro-strategy applied by the interpreter when dealing with constraints such as syntactic asymmetry between source and target languages. In spite of small local differences between and among them, all three "safe" strategies (i.e., waiting, stalling and chunking) require considerably more cognitive processing resources than baseline. Consequently, one could argue that the amount of cognitive load experienced by interpreters might be causally related to the amount of restructuring they engage in.

Third, the local fluctuations of cognitive load reflected in the model seem to be of a magnitude that does not lend support to Gile's "tightrope hypothesis," according to which "most of the time, interpreters work near saturation level" (Gile 1999). In fact, even if we assume that the local (and thus relative) maximum load experienced during any of the four strategies represents the absolute maximum load, interpreters still work below saturation levels a considerable part of the time. If, on the other hand, the local maximum load does not represent the absolute maximum load, we would have to assume that interpreters often work well under saturation levels *even* when simultaneously interpreting verb-final structures into a verb-initial language.

On the whole, the analytical approach presented here shows that sufficient evidence exists in neighboring disciplines for us to start explainimg the processes at work during SI between structurally asymmetrical language pairs, notably from SOV to SVO structures. Based on this evidence, a solid theoretical case can be made within an information processing paradigm to suggest that simultaneous interpreting of SOV into SVO structures generates more cognitive load than interpreting SVO into SVO structures.

Notes

[1] Whereas in Broadbent's (1958) and Treisman's (1969) models the bottleneck is situated at the level of perception (entailing early selection), Deutsch and Deutsch (1963) and Norman (1968) believe that this bottleneck is at the response level, arguing for late selection.

2. The notion of chunks in cognitive psychology, referring to individual units of information, need not correspond to that used in linguistics, consisting of a single content word surrounded by a number of function words (Abney 1991).

3. Minimal attachment: to attach the new constituent using the minimum number of nodes.

4. Late closure: to attach to the current verb phrase when there is more than one minimal attachment.

5. Pio (2002) compares speeches presented at 108 wpm and 145 wpm, whereas Seeber (2005) uses materials recorded at 120 wpm and 145 wpm respectively.

References

Alexieva, B. (1998). Consecutive interpreting as a decision process. In A. Beylard-Ozeroff, J. Kralova & B. Moser-Mercer (Eds.), *Translators' strategies and creativity*. Amsterdam/Philadelphia: John Benjamins, 181–188.

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

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

Bader, M. & Lasser, I. (1994). German verb-final clauses and sentence processing: Evidence for immediate attachment. In C. Clifton, L. Frazier & K. Rayner (Eds.), *Perspectives in sentence processing*. Hillsdale, NJ: Erlbaum, 225–242.

Barik, H. C. (1973). Simultaneous interpretation: Temporal and quantitative data. *Language and Speech* 16, 237–270.

Boland, J. E., Tanenhaus, M. K. & Garnsey, S. M. (1990). Evidence for the immediate use of verb control information in sentence processing. *Journal of Memory and Language* 29, 413–432.

Broadbent, D. E. (1958). *Perception and communication*. London: Pergamon Press.

Chase, W. G. & Simon, H. A. (1973). The mind's eye in chess. In W. G. Chase (Ed.), *Visual information processing*. New York: Academic Press, 215–281.

Chernov, G. V. (1992). Conference interpretation in the USSR: History, theory, new frontiers. *Meta* 37 (1), 149–162.

Chomsky, N. & Miller, G. A. (1963). Introduction to the formal analysis of natural languages. In R. D. Luce, R. R. Bush & E. Galanter (Eds.), *Handbook of mathematical psychology*. Vol. 2. New York: Wiley, 269–321.

Christoffels, I. K. (2004). *Cognitive studies in simultaneous interpreting*. Enschede: PrintPartners Ipskamp.

Clark, H. H. (1975). Bridging. In R. C. Schank & B. L. Nash-Webber (Eds.), *Theoretical issues in natural language processing*. New York: Association for Computing Machinery, 169–174.

Coltheart, M. (1972). Visual information-processing. In P. C. Dodwell (Ed.), *New horizons in psychology*. Vol. 2. Harmondsworth, UK: Penguin, 62–85.

Cowan, N. (2000). Processing limits of selective attention and working memory: Potential implications for interpreting. *Interpreting* 5 (2), 117–146.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, 24 (1), 87–185.

Cowan, N. (2005). Working-memory capacity limits in a theoretical context. In C. Izawa & N. Ohta (Eds.), *Human learning and memory: Advances in theory and application. The 4th Tsukuba International Conference on Memory.* Mahwah, NJ: Erlbaum, 155–175.

Cowan, N., Chen, Z. & Rouder, J. N. (2004). Constant capacity in an immediate serial-recall task: A logical sequel to Miller (1956). *Psychological Science* 15, 634–640.

Craik, F. I. M. & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior* 11, 671–684.

Dawrant, A. (1996). *Word order in Chinese-English simultaneous interpretation: An initial exploration*. Unpublished MA thesis, Fu Jen University.

De Groot, A. D. (1965). *Thought and choice in chess*. The Hague: Mouton.

De Groot, A. M. B. (2000). A complex-skill approach to translation and interpreting. In S. Tirkkonen-Condit & R. Jääskeläinen (Eds.), *Tapping and mapping the processes of translation and interpreting*. Amsterdam: John Benjamins, 53–68.

Ericsson, K. A. & Kintsch, W. (1995). Long-term working memory. *Psychological Review* 102 (2), 211–245.

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*. Amsterdam: John Benjamins, 273–317.

Fagot, C. & Pashler, H. (1992). Making two responses to a single object: Implications for the central attentional bottleneck. *Journal of Experimental Psychology: Human Perception and Performance* 18, 1058–1079.

Ferreira, F., Baily, K. D. G. & Ferraro, V. (2002). Good-enough representations in language comprehension. *Current Directions in Psychological Science* 11 (1), 11–15.

Fiebach, C. J., Schlesewsky, M. & Friederici A. D. (2001). Syntactic working memory and the establishment of filler-gap dependencies: Insights from ERPs and fMRI. *Journal of Psycholinguistic Research* 30 (3), 321–338.

Fodor, J. A., Bever, T. G. & Garrett, M. F. (1974). *The psychology of language*. New York: McGraw-Hill.

Frazier, L. (1987). Sentence processing: A tutorial review. In M. Coltheart (Ed.), *Attention and performance 12: The psychology of reading*. Hove, UK: Lawrence Erlbaum, 559–586.

Friederici, A. D. & Bornkessel, I. (2003). Missing the syntactic piece. Commentary to Ruchkin et al. *Behavioral and Brain Sciences* 26 (6), 735–736.

Garnham, A. (1987). *Mental models as representation of discourse and text*. Chichester, UK: Horwood.

Garrod, S., O'Brien, E. J., Morris, R. K. & Rayner, K. (1990). Elaborative inferencing as an active or passive process, *Journal of Experimental Psychology: Learning, Memory, and Cognition* 16 (2), 250–257.

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

Gibson, E. (1998). Linguistic complexity: Locality of syntactic dependencies. *Cognition* 68, 1–76.

Gibson, E. (2000). The dependency locality theory: A distance-based theory of linguistic complexity. In: Y. Miyashita, A. Marantz & W. O'Neil (Eds.), *Image, language, brain*. Cambridge, MA: MIT Press, 95–126.

Gibson, E. & Pearlmutter, N. (1998). Constraints on sentence comprehension. *Trends in Cognitive Sciences* 2, 262–268.

Gile, D. (1992). Predictable sentence endings in Japanese and conference interpretation. *The Interpreters' Newsletter*, Special Issue 1, 12–24.

Gile, D. (1995). *Regards sur la recherche en interprétation de conférence*. Lille: Presses universitaires de Lille.

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

Gile, D. (1999). Testing the Effort Models' tightrope hypothesis in simultaneous interpreting: A contribution. *Hermes* 23, 153–171.

Gile, D. (2008). Local cognitive load in simultaneous interpreting and its implications for empirical research. *Forum* 6 (2), 59–77.

Gobet, F. & Simon, H. A. (1996a). Recall of random and distorted positions: Implications for the theory of expertise. *Memory and Cognition* 24, 493–503.

Gobet, F. & Simon, H. A. (1996b). Recall of rapidly presented random chess positions is a function of skill. *Psychonomic Bulletin and Review* 3, 159–163.

Gobet, F. & Simon, H. A. (1998). Expert chess memory: Revisiting the chunking hypothesis. *Memory* 6, 225–255.

Goldman-Eisler, F. (1972). Segmentation of input in simultaneous translation. *Journal of Psycholinguistic Research* 1 (2), 127–140.

Gordon, P. C., Hendrick, R. & Johnson, M. (2001). Memory interference during language processing. *Journal of Experimental psychology: Learning, Memory and Cognition* 27, 1411–1423.

Grodner, D. & Gibson, E. (2005). Consequences of the serial nature of linguistic input. *Cognitive Science* 29, 261–291.

Harley, T. A. (2001). *The psychology of language*. 2nd edition. Hove: Psychology Press.

Harris, R. J. & Monaco, G. E. (1978). Psychology of pragmatic implication: Information processing between the lines. *Journal of Experimental Psychology: General* 107, 1–22.

Hawkins, J. A. (1994). *A performance theory of order and constituency*. Cambridge: Cambridge University Press.

Heil, M., Rolke, B. & Pecchineda, A. (2004). Automatic semantic activation is no myth. *Psychological Science* 15 (12), 852–856.

Hemforth, B., Konieczny, L. & Strube, G. (1993). Incremental syntax processing and parsing strategies. In *Proceedings of the 15th Annual Conference on the Cognitive Science Society, July, 1993*. Hillsdale, NJ: Erlbaum, 539–545.

Hyönä, J., Tommola, J. & Alaja, A. (1995). Pupil dilation as a measure of processing load in simultaneous interpreting and other language tasks. *The Quarterly Journal of Experimental Psychology* 48A (3), 598–612.

Ilg, G. (1959). *L'enseignement de l'interprétation à l'école d'Interprètes de L'université de Genève*. Genève: Université de Genève.

Ilg, G. (1978). L'apprentissage de l'interprétation simultanée. *Parallèles* 1, 69–99.

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

Johnson-Laird, P. N. (1983). *Mental models*. Cambridge: Cambridge University Press.

Just, M. A. & Carpenter, P. A. (1992) A capacity theory of comprehension: Individual differences in working memory. *Psychological Review* 99 (1), 122–149.

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

Kamide, Y. & Mitchell, D. C. (1999). Incremental pre-head attachment in Japanese parsing. *Language and Cognitive Processes* 14 (5/5), 631–662.

Kamide, Y., Scheepers, C. & Altmann, G. T. M. (2003). Integration of syntactic and semantic information in predictive processing: Cross-linguistic evidence from German and English. *Journal of Psycholinguistic Research* 32 (1), 37–55.

Kimball, J. (1973). Seven principles of surface structure parsing in natural language. *Cognition* 2, 15–47.

Kintsch, W. & Van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological Review* 85, 363–394.

Kintsch, W., Patel, V. L. & Ericsson, K. A. (1999). The role of long-term working memory in text comprehension. *Psychologia* 42, 186–198.

Kirchhoff, H. (1976). Das Simultandolmetschen: Interdependenz der Variablen im Dolmetschprozess, Dolmetschmodelle und Dolmetschstrategien. In H. W. Drescher & S. Scheffzek (Eds.), *Theorie und Praxis des Übersetzens und Dolmetschens*. Frankfurt: Lang, 59–71.

Knowles, W. B. (1963). Operator loading tasks. *Human Factors* 5 (2), 155–161.

Kohn, K. & Kalina, S. (1996). The strategic dimension of interpreting. *Meta* 41 (1), 119–138.

Konieczny, L. (1996). *Human sentence processing: A semantics-oriented parsing approach*. Doctoral dissertation. University of Freiburg: IIG-Berichte 3/96.

Konieczny, L. & Döring, P. (2003). Anticipation of clause-final heads. Evidence from eye-tracking and SRNs. In *Proceedings of the 4th ICCS/ASCS Joint International Conference on Cognitive Science. Sydney, Australia*, 330–335.

Konieczny, L. & Hemforth, B. (1994). Incremental parsing with lexicalized grammars. In G. Strube (Ed.), *Current research in cognitive science at the Center for Cognitive Science*. Universität Freiburg: IIG- Berichte 1/94, 33–54.

Lederer, M. (1978). Simultaneous interpretation — units of meaning and other features. In D. Gerver & H. W. Sinaiko (Eds.), *Language interpretation and communication*. New York: Plenum Press, 323–332.

Lederer, M. (1981). *La traduction simultanée: expérience et théorie*. Paris: Minard.

Lee, T.-H. (2002). Ear voice span in English into Korean simultaneous interpretation. *Meta* 47 (4), 596–606.

Lewis, R. L. (1996). Interference in short-term memory: The magical number two (or three) in sentence processing. *Journal of Psycholinguistic Research* 25 (1), 93–115.

Lonsdale, D. (1997). Modeling cognition in SI: Methodological issues. *Interpreting* 2 (1/2), 91–117.

MacDonald, M. C. & Christiansen M. H. (2002) Reassessing working memory: Comment on Just and Carpenter (1992) and Waters and Caplan (1996), *Psychological Review* 109 (1), 35–54.

Massaro D. W. & Shlesinger, M. (1997) Information processing and a computational approach to the study of simultaneous interpretation. *Interpreting* 1 (1/2), 13–53.

Mayer, L. V. (1988). *Voice and diction*. 8th edition. Dubuque, IA: Brown.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review* 63, 81–97.

Miller, G. A. (1989). George A. Miller. In L. Gardner, (Ed.), *A history of psychology in autobiography*. Vol. VIII. Stanford, CA: Stanford University Press, 391–418.

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

Moser-Mercer, B. (1994). Paradigms gained or the art of productive disagreement. In S. Lambert & B. Moser-Mercer (Eds.), *Bridging the gap: Empirical research in simultaneous interpretation*. Amsterdam/Philadelphia: John Benjamins, 17–24.

Moser-Mercer, B. (1997). Beyond curiosity: Can interpreting research meet the challenge? In J. H. Danks, S. B. Fountain, M. K. McBeath & G. M. Shreve (Eds.), *Cognitive processes in translation and interpreting*. Thousand Oaks, CA: Sage Publishing, 176–195.

Oléron, P. & Nanpon, H. (1965). Recherches sur la traduction simultanée. *Journal de psychologie normale et pathologique* 62, 73–94.

Osterhout, L. (1994). Event-related brain potentials as tools for comprehending language comprehension. In C. Clifton, Jr., L. Frazier & K. Rayner (Eds.), *Perspectives on sentence processing*. Hillsdale, NJ: Erlbaum, 15–44.

Paradis, M. (1994). Toward a neurolinguistic theory of simultaneous translation: The framework. *International Journal of Psycholinguistics* 9 (2), 133–145.

Pashler, H. & Johnston, J. C. (1998). Attentional limitations in dual-task performance. In H. Pashler (Ed.), *Attention*. Hove, UK: Taylor & Francis, 155–189.

Pio, S. (2003) The relation between ST delivery rate and quality in simultaneous interpretation. *The Interpreters' Newsletter* 12, 69–100.

Ramshaw, L. A. & Marcus, M. P. (1995). Text chunking using transformation-based learning. In D. Yarowsky & K. Church (Eds.), *Proceedings of the Third Workshop on Very Large Corpora*. Cambridge, MA: Massachusetts Institute of Technology, 82–94.

Riccardi, A. (1998). Interpreting strategies and creativity. In A. Beylard-Ozeroff, J. Kralova, & B. Moser-Mercer (Eds.), *Translators' strategies and creativity*. Amsterdam/Philadelphia: John Benjamins, 171–179.

Riccardi, A. & Snelling, C. (1997). Sintassi tedesca: vero o falso problema per l'interpretazione? In L. Gran & A. Riccardi (Eds.), *Nuovi orientamenti negli studi sull'interpretazione*. Padova: CLEUP, 143–158.

Rinne, J. O., Tommola, J., Laine, M., Krause, B. J., Schmidt, D., Kaasinen, V., Teräs, M., Sipilä, H. & Sunnari, M. (2000). The translating brain: Cerebral activation patterns during simultaneous interpreting. *Neuroscience Letters* 294, 85–88.

Ruchkin, D. S., Grafman, J., Cameron, K. & Berndt, R. S. (2003). Working memory retention systems: A state of activated long-term memory. *Behavioral and Brain sciences* 26 (6), 709–777.

Sanders, A. F. (1979). Some remarks on mental load. In N. Moray (Ed.), *Mental workload: Its theory and measurement*. New York: Plenum Press, 41–77.

Scheepers, C., Hemforth, B. & Konieczny, L. (1999). Incremental processing of German verb-final constructions: Predicting the verb's minimum (!) valency. In *Proceedings of the Second International Conference on Cognitive Science (ICCS/JCSS99), Tokyo, July 27–30, 1999* (no pagination).

Schneider, W., Dumais, S. T. & Shiffirn, R. M. (1984) Automatic and control processing and attention. In R. Parasuraman & D. R. Davies (Eds.), *Varieties of attention*. New York: Academic Press, 1–27.

Schumacher, E. H., Saymour, 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.

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

Seeber, K. G. (2005). Temporale Aspekte der Antizipation beim Simultandolmetschen von SOV-Strukturen aus dem Deutschen. In A. Künzli (Ed.), *Empirical research into translation and interpreting: processes and products*. Neuchâtel: Institut de linguistique de l'Université de Neuchâtel (*Bulletin Suisse de linguistique appliquée Vals-Alsa* 81), 123–140.

Seeber, K. G. (2007). Thinking outside the cube: Modeling language processing tasks in a multiple resource paradigm. In *Conference proceedings. Interspeech 2007, Antwerp, Belgium*, 1382–1385.

Seleskovitch, D. (1978). *Interpreting for international conferences*. Washington DC: Pen and Booth.

Seleskovitch, D. (1984). Les anticipations de la compréhension. In D. Seleskovitch & M. Lederer, *Interpréter pour traduire*. Paris: Didier Erudition, 273–283.

Setton, R. (1999). *Simultaneous interpretation: A cognitive-pragmatic analysis*. Amsterdam/Philadelphia: John Benjamins.

Setton, R. & Motta, M. (2007). Syntacrobatics: Quality and reformulation in simultaneous-with-text. *Interpreting* 9 (2), 199–230.

Shlesinger, M. (2000). *Strategic allocation of working memory and other attentional resources in simultaneous interpreting*. Unpublished doctoral dissertation, Bar Ilan University.

Simon, H. A. (1974). How big is a chunk? *Science* 183, 482–488.

Singer, M. (1994). Discourse inference process. In M. A. Gernsbacher (Ed), *Handbook of psycholinguistics*. San Diego: Academic Press, 479–515.

Slak, S. (1970). Phonemic recoding of digital information. *Journal of Experimental Psychology* 86, 398–406.

Stabler, E. P. (1994). The finite connectivity of linguistic structures. In C. Clifton Jr., L. Frazier & K. Rayner (Eds.), *Perspectives on sentence processing*. Hillsdale, NJ: Erlbaum, 303–336.

Styles, E. A. (1997). *The psychology of attention*. Hove, UK: Psychology Press.

Trueswell, J. C., Tanenhaus, M. K. & Kello, C. (1993) Verb-specific constraints in sentence processing: Separating effects of lexical preference from garden paths. *Journal of Experimental Psychology: Learning, Memory and Cognition* 19, 528–553.

Ueno, M. & Polinsky, M. (2005). Maximizing processing in an SOV language (Manuscript). http://internal.psychology.illinois.edu/~ueno/UenoCorpus.pdf (accessed 23 December 2010).

Van Dam, I. M. (1989). Strategies of simultaneous interpretation. In L. Gran & J. Dodds (Eds.), *The theoretical and practical aspects of teaching conference interpretation*. Udine: Companotto Editore, 167–176.

Van den Broek, P. (1994). Comprehension and memory of narrative texts: Inferences and coherence. In M. A. Gernsbacher (Ed.), *Handbook of psycholinguistics*. San Diego: Academic Press, 539–588.

Van Dijk, T. A. & Kintsch, W. (1983). *Strategies of discourse comprehension*. New York: Academic Press.

Welford, A. T. (1952). The "psychological refractory period" and the timing of high speed performance — a review and a theory. *British Journal of Psychology* 43, 2–19.

Wickens, C. D. (1976). The effects of divided attention on information processing in manual tracking. *Journal of Experimental Psychology: Human Perception and Performance* 2, 1–13.

Wickens, C. D. (1984). Processing resources in attention. In R. Parasuraman & D. R. Davies (Eds.), *Varieties of attention*. New York: Academic Press, 63–102.

Wickens, C. D. (2002). Multiple resources and performance prediction. *Theoretical Issues in Ergonomics Science* 3 (2), 159–177.

Wilson, D. (1999). Relevance and relevance theory. In R. Wilson & F. Keil (Eds.), *MIT encyclopedia of the cognitive sciences*. Cambridge: MIT Press, 719–722.

Wilss, W. (1978). Syntactic anticipation in German-English simultaneous interpreting. In D. Gerver & H. W. Sinaiko (Eds.), *Language interpretation and communication*. New York: Plenum Press, 343–352.

Yngve, V. H. (1960). A model and hypothesis for language structure. *Proceedings of the American Philosophical Society* 104, 444–466.

Zanetti, R. (1999). Relevance of anticipation and possible strategies in the simultaneous interpretation from English into Italian. *The Interpreters' Newsletter* 9, 79–98.

Zwaan, R. A. (1999). Situation models: The mental leap into imaged worlds. *Current Directions in Psychological Science* 8 (1), 15–18.

Author's address

Kilian Seeber
Ecole de Traduction et d'Interprétation
Université de Genève
40, boulevard du Pont-d'Arve
CH-1211 Genève 4
Switzerland

kilian.seeber@unige.ch

About the author

Kilian Seeber is Assistant Professor at ETI's Interpreting Department (University of Geneva). He completed his undergraduate training in translation and interpreting at the University of Vienna, did his graduate work in interpreting at the University of Geneva and his post-doctoral research in psycholinguistics at the University of York. The main focus of his research to date has been on cognitive aspects of language processing, more specifically on anticipation and working memory in simultaneous interpreting.