who had accepted to participate in the experiment, it is unlikely that the reason was technical, perhaps the poor quality of the recording at these points in the speech – otherwise, more interpreters would probably have had problems with the same segments. A third point is that some interpreters who interpreted the speech twice in a row (after a short break of a few minutes) made some errors the second time in segments which they had interpreted correctly the first time. This is an intriguing point: if they had overcome a difficulty when first interpreting a speech, why should they not be able to overcome it when interpreting it a second time, with the advantage of having become familiar with it and having had time to think about it? One possible explanation would be fatigue. However, the speech was less than 11 minutes long, and subjects were allowed to rest for a couple of minutes before starting to reinterpret it. In view of the fact that when interpreting in the field, the same subjects take turns of 30 minutes in the simultaneous interpreting booth, fatigue is also an unlikely explanation for the phenomenon. (Gile 1999a is a full report on this experiment – see also Section 9, which discusses this experiment again around the concept of 'Tightrope Hypothesis').

Observations of errors made by professionals in speech segments containing no apparent obstacle are intriguing, and trying to understand the reasons behind them seemed important, if only in order to help students understand why interpreting is so difficult and accept this as a fact of life rather than as a worrying sign of incompetence. Insight into the mechanisms leading to performance flaws was also sought in the hope of finding ideas and methods to help overcome the obstacles.

In this mindset, I developed an Effort Model for simultaneous interpreting, which was first sketched out in a paper on the relative difficulty of interpreting as a function of the specific pairs of languages involved (Gile 1983b). Ever since, I have been developing it and extending the analysis. This chapter is an up-to-date discussion of the Models (there are now several), which are central in my teaching of interpreting and have been adopted as a conceptual framework by many interpreting teachers – and, as it turned out, by researchers (see Gile 2008) – over the past 25 years. The Effort Models for simultaneous interpreting, for consecutive interpreting and for sight translation are introduced here as a simple set of constructs for explanatory purposes. Since in the literature, they are also discussed in the context of considerations from cognitive psychology, I have added some clarifications on how they relate to and differ from models and theories from mainstream cognitive psychology.

2. Automatic operations, processing capacity and interpreting Efforts

2.1 Automatic and non-automatic operations

The development of the Models originated in two intuitive ideas based on observation and introspection:

– Interpreting requires some sort of 'mental energy' that is only available in limited supply.
– Interpreting takes up almost all of this mental energy, and sometimes requires more than is available, at which times performance deteriorates.

The idea that there is some association between the deterioration of the interpreter's performance and some kind of overload was not new. It had already been mentioned by Pinter (1969), as well as by several other authors, mainly in the context of discussions of the role of short-term memory in simultaneous interpreting (Fukuii & Asano 1961; Kade & Cartellieri 1971; Lederer 1978; Moser 1978; Wilss 1978). Subsequent reading in cognitive psychology provided useful information revolving around the concepts of *attention* and 'automatic' and 'non-automatic' operations, thus establishing a link between intuitive ideas from interpreting practice and some theoretical and empirical research.

Toward the end of the 1940s, Claude Shannon, an engineer working on communication, formulated the idea that any channel serving to transmit information had a finite transmission capacity beyond which information losses occurred (Shannon 1948). This idea, which had come out of studies on electric communication lines, was taken up by cognitive psychologists who adapted it to the case of the human mind (Broadbent 1958; Moray 1967; Kahneman 1973; Norman 1976). The idea is that some mental operations ('non-automatic operations') require attention (alternative names are 'attentional resources' and 'processing capacity'), and others ('automatic operations') do not. Such non-automatic operations also take time, whereas automatic operations are very fast. Non-automatic operations take processing capacity from a limited available supply (whether all of them take it from the same single reservoir or not is under debate). When the processing capacity available for a particular task is insufficient, performance deteriorates.

The distinction between automatic and non-automatic operations is sometimes difficult to make, as non-automatic operations vary in the processing capacity they require and may become automatic after enough repetition (see for instance Eysenck & Keane 1990). Gradual automation of cognitive operations is important in interpreting skills acquisition and will be discussed further in the next chapters.

Finding these concepts and theories in cognitive psychology was encouraging: if operations involved in interpreting were non-automatic, there was some basis for constructing an interpreting model around the ideas of processing capacity requirements and processing capacity limitations with prospects for good explanatory power as regards various phenomena experienced and observed while interpreting.

According to cognitive psychology, non-automatic operations are those which cannot be or are not automated, such as detecting a brief stimulus, identifying a non-familiar stimulus or a familiar stimulus presented under poor conditions, storing information in memory for later use, preparing for a non-automated response, controlling

the accuracy of a movement or manipulating symbols in a person's cognitive system. Automatic operations include decoding a familiar stimulus presented under favourable conditions, triggering an automated response and operating a motor programme without control (Richard 1980: 149–150). Again, this distinction is a simplification of reality, if only because it may sometimes be difficult to discriminate between long and short stimuli or between familiar and unfamiliar stimuli. Nevertheless, it is shown in the next sections of this chapter that the operations making up interpreting as defined below clearly include components on the non-automatic side.

2.2 Interpreting Efforts

Drawing on my experience as a practitioner of simultaneous interpreting and as an instructor, I thought of attempting to analyze its operation with conceptual entities immediately familiar to interpreters and to students, namely a listening and analysis component, a speech production component, and a short-term memory component. I called these components 'Efforts' to stress their effortful nature, as they include deliberate action which requires decisions and resources.

2.2.1 *The Listening and analysis Effort*

The Listening and analysis Effort (or 'Listening Effort' for short) was defined as consisting of all comprehension-oriented operations, from the subconscious analysis of the sound waves carrying the source-language speech which reach the interpreter's ears through the identification of words to the final decisions about the 'meaning' of the utterance. In signed language interpreting, a parallel Viewing and Analysis Effort can be defined when the interpreter works from a signed language into a spoken language.

It is not yet clear how far the analysis of the meaning of the source-language speech must go before interpreting is possible. In the discussion of the interpreting and translation of specialized speeches and texts in Chapter 4, it is suggested that such comprehension goes at least as far as understanding the general underlying logic of each sentence. Even by the most conservative standards, one can say that except for some names which interpreters may simply try to imitate phonetically, interpreting requires at the very least the recognition of words in the source-language speech. This is enough to put the Listening and analysis component in the non-automatic category. The reason is that there is no one-to-one relation between the sound reaching one's ears and any single phoneme, word, or group of words pronounced by a speaker (the same applies to the visual flows perceived by one's eyes when interpreting from signed languages – a view which signed-language specialists Carol Patrie and Robert E. Johnson of Gallaudet University confirmed in a personal exchange). There is some variability in the way such words are pronounced, not only from one individual to another, but also in the same speaker repeating the same speech segment.

This is also why it is very difficult to recognize words on sonograms (graphic representations of sound). According to Guibert (1979), even human experts familiar with phonetics and with the lexical and syntactic rules of a language never manage to read correctly more than 75% of the phonetic segments on a sonogram. It is therefore not surprising that machines, which do not have at their disposal as much lexical and syntactic information, nor a level of knowledge of the world comparable to a human's, are unable to recognize natural chained (continuous) speech (as opposed to speech in which there are pauses between the words) with 100 percent reliability. This limitation persists even when they have gone through a 'learning process' with the voice of individual speakers (a necessary process when using dictation software).

In order for words to be recognized, acoustic features of the incoming sounds have to be analyzed and compared with patterns stored in the listener's long-term memory (or in the hard disk of a computer, in the case of speech recognition software). Following a complicated process involving knowledge of the probabilistic structure of the relevant language, the context and the situation, the listener decides that particular sound sequences correspond to particular words (see for example Hörmann 1972; Clark & Clark 1977; Costermans 1980; Noizet 1980; Matthei & Roeper 1985; Greene 1986). Undoubtedly, speech recognition as it occurs in interpreting has non-automatic components.

Actually, interpreters know that interpreting involves much more than speech recognition. Some kind of semantic representation of the content of source speeches is always present, which includes plausibility analyses (see Chapter 4, 5 and 6) and probably some anticipation. Chernov (1973) conducted an experiment in which he made students interpret sentences that seemed to be leading in a certain direction and then veered off to an unexpected ending. He found they were usually interpreted according to the direction they were taking initially, not as they finally turned out to end. His subjects not only identified words, but also made inferences about their meaning and anticipated on-line. Several studies by Gerver (1976), Lambert (1988) and others focused on comparisons of recall and recognition under various conditions and also led to results suggesting that when interpreting simultaneously, interpreters do achieve a level of comprehension much beyond the recognition of individual words.

No sophisticated research is necessary to ascertain that interpreting comprehension goes beyond word recognition: field observation provides ample evidence for the fact in practically any interpreted speech. As an illustration, below are French renditions by five professional interpreters of the same English speech segment. The material is taken from one experiment (Gile 1999b) involving in-laboratory interpreting of the same recordings of actual conference speeches.

Original English sentence (by speaker):

> "I was hoping to encourage the oil people to give a little bit back to the countries that they take the oil from"

French versions produced by five professional interpreters:

> «*Je voulais encourager les pétroliers à rendre un peu de leur butin aux pays où ils vont le prélever* » (Interpreter A)
>
> « *Je pensais pouvoir encourager les compagnies pétrolières à restituer un peu des profits aux pays qu'ils exploitent* » (Interpreter B)
>
> « *J'espérais que les compagnies pétrolières rendraient quelque chose à ceux dont ils prennent le pétrole* » (Interpreter C)
>
> « *Je voulais encourager les pétroliers entre guillemets à rendre un petit peu de ce qu'ils ponctionnent aux pays qui ont du pétrole* » (Interpreter D)
>
> « *… pour en quelque sorte sensibiliser les gens du secteur pétrolier afin qu'ils rendent un petit peu de ce qu'ils gagnent aux pays où ils prennent le pétrole* » (Interpreter E)

In this example, "oil people" was interpreted into "*pétroliers*", "*compagnies pétrolières*", "*gens qui font de la recherche pétrolière*" and "*gens du secteur pétrolier*". None of these rendering are word-for-word translations. In particular, *compagnies pétrolières* means 'oil companies' and shows that the English term was interpreted instead of being taken literally. As to the rest of the sentence, the word "*butin*" ('loot') used by interpreter A, the expression "*les profits*" used by interpreter B, "*sensibiliser*" ('raise the awareness') show the interpreters' understanding of the economic meaning of the situation and of the moral stance taken by the speaker. Many similar examples of the interpretive nature of comprehension during conference interpreting have been given by other authors, in particular by Seleskovitch and Lederer of ESIT. Such observations only confirm the difference between human interpreters and translators on the one hand, and machine translation programmes based on linguistic analysis on the other. The latter's failure can be attributed to their present inability to relate linguistic signs to knowledge of the world so as to disambiguate and solve other problems arising from the author's linguistic errors, mistakes in substance, and deviations from standard language and logic.

One might also add that the comprehension effort is probably more intense for interpreters than it is for conference delegates, i.e. the people to whom the speaker is talking (see also Chapter 8 and Chapter 9):

- While they are interpreting, interpreters have to concentrate on everything the speaker says whereas delegates can select the information they are interested in.
- The interpreters' relevant extralinguistic knowledge, and often the terminological part of their linguistic knowledge, are less comprehensive than the delegates'.

It follows that comprehension during interpreting is a non-automatic process. As will be explained later in this chapter, it is also critical in terms of processing capacity management.

2.2.2 *The Production Effort*

This is the name given here to the 'output part' of interpreting. In simultaneous interpreting, it can be defined as the set of operations extending from the mental representation of the message to be delivered to speech planning and the performance of the speech plan, including self-monitoring and self-correction when necessary.

As in speech comprehension, the impression of effortlessness in speech production is deceptive. Matthei and Roeper stress that (1985: 114),

> … the fact that virtually all people make many false starts, add ums and ahs, and often speak ungrammatically, suggests that production may be making quite a number of very substantial demands on our linguistic systems.

According to Holmes (1988: 324), "Speakers' efforts to realize their intentions are seldom completely fluent and error-free." For Clark and Clark (1977: 226), "Speaking is problem solving." Such problems become particularly salient in hesitations associated with the search for lexical units and with syntactic decision-making (Maclay & Osgood 1959): it often takes time to find the right word, and it often takes time to decide how to steer the sentence at a syntactic junction. Hesitations are the main symptom that makes speakers and their listeners aware of speech production difficulties. Hesitations are also the main factor that determines effective speech rate, i.e. the number of words actually uttered per unit of time, as opposed to the more mechanical articulation rate (Goldman-Eisler 1958; Clark & Clark 1977; Costermans 1980).

Speech production problems account for a number of interesting phenomena. One of them is that speakers tend to "be tempted and constrained to having recourse to ready made verbal sentences, phrases and clichés" (Goldman-Eisler 1958: 67–68). Cherry (1978: 79) explains that "We become prone to verbal habits. It is only too easy to use clichés, proverbs and slogans as a substitute for reasoned statements." Goldman-Eisler found that "Fluent speech was shown to consist of habitual combinations of words such as were shared by the language community and such as had become more or less automatic" (1958: 67), and concludes that meaning itself may be "guided through these channels and modified as a result" (1958: 68).

These observations suggest one reason why speech production under interpreting conditions may be difficult. People who speak on their own behalf are free to speak their own mind and bypass possible production difficulties by rearranging the sequence of information and ideas, or by dropping or modifying some of these or using standard phrases which are not necessarily quite in line with their initial message. In contrast, interpreters find themselves forced to follow rather closely the path chosen by another speaker, if only because waiting for a sentence to finish before starting to interpret it would cause excessive short-term memory load (see the discussion of memory load later in this chapter). In addition, "habitual combinations of words" generally differ from language to language, which makes the interpreter's speech production task more

difficult than the speaker's – at least in this respect (see also the discussion of speech production in Chapter 9).

On the other hand, the very fact that lexical and syntactic choices are made by the speaker can in some cases help the interpreter, if s/he can make similar or 'parallel' choices in the target language, or at least use them in some way when retrieving words from his/her mental lexicon and making his/her own syntactic choices. Such verbal piggy-back riding is often done, at times consciously, and seems to help interpreters speak faster than they can when they have only their natural fluency to rely on – but entails risks, as explained below.

If an interpreter uses source-language words and structures to construct his/her own target-language speech, the speech production process becomes more vulnerable:

– Following the source-language structure and lexical choices in one's target-language speech is risky because the interpreter may get stuck because of syntactic and grammatical differences between the languages.

In a training session in the classroom, one student tried to interpret into French the following English sentence by following its structure and the speaker's lexical choices:

> "This movement was shown by our team as consisting of three parts …"

The student started with: "*Ce mouvement a été montré par notre équipe …*", which mirrors the English sentence structure, and could not construct a natural, grammatical segment to complete the sentence, as French does not allow the passive of the verb *montrer* (to show) to be followed by a structure similar to "as consisting of." A natural translation of the sentence into French would have put "team" in the beginning of the target version, in something like:

> *Notre équipe a montré que ce mouvement se compose de trois parties …*

("Our team showed that this movement consists of three parts")

– Besides the risk of getting stuck, when following the target-language structure and lexicon, interpreters may find themselves deprived of part of their own favourite productive linguistic resources as speakers (words and structures) which they might put to use if they were to produce a speech on the basis of meaning rather than on the basis of a linguistic structure borrowed from another speaker, in another language at that.
– Third, such transcoding is associated with great danger of linguistic interference between the two languages, be it gross interference resulting in grammatical errors, mispronunciations and false cognates, or more discrete interference that will make the interpreter's speech more hesitant, less idiomatic, less clear, less pleasant to listen to.

– Fourth, by focusing on language, the interpreter is in greater danger of processing the incoming speech more superficially than if s/he produced the speech from the meaning. This may generate more errors, because the interpreter will do less plausibility-testing (see Chapter 5).

For all these reasons, which they do not necessarily explain in so many words, interpreting instructors tend to say that the rule to be followed whenever possible is to produce the target-language speech on the basis of the meaning, not the words of the source-language speech (but see Chapter 9 on the role of Translinguistic Equivalences).

The fact that interpreting constraints force interpreters to deviate from their habitual speech production patterns may account for the poor quality of language output in students' interpreting exercises. In a naturalistic study conducted on five native speakers of French at ESIT, Paris, during a whole academic year (Gile 1987), the number of deviations from acceptable linguistic standards (as indicated by native informants) per sequence of 100 words was measured in three types of exercises: presentations by the students, exercises in consecutive and exercises in simultaneous interpreting. The target language was French in all cases. Deviations were found to be more numerous in consecutive than in presentations, and more numerous in simultaneous than in consecutive. Interestingly, most of these deviations did not seem to be due to interference between source and target language, at least not in any direct, visible way. No systematic comparison was done with the output quality of professional interpreters, but scrutiny of consecutive and simultaneous interpreting transcripts from other experiments seems to suggest that overall, linguistic deviation rates in professionals are much lower (however, see Alonso Bacigalupe 2006 for different findings). It appears that the effects of interpreting constraints on production are stronger in simultaneous than in consecutive, probably because of differences between the two modes, both in processing capacity management and in time constraints – the two are obviously linked.

A further difficulty, already mentioned for the Comprehension Effort in Section 2.1, arises from the fact that interpreters often have to produce speech in fields with which they are not necessarily familiar. Neither are they always familiar with the particular sociolect used by the relevant groups with respect to words, particularly technical terms, and phraseology – the same issue was highlighted in previous chapters when discussing written translation. Again, speech production in interpreting is clearly a non-automatic operation.

2.2.3 *The Memory Effort*

During interpreting, short-term memory operations (up to a few seconds) succeed each other without interruption. Some are due to the lag between the moment speech sounds are heard and the moment they are interpreted: phonetic segments may have to be added up in memory and analyzed until they allow identification of a word or phoneme.

To take only one example, when spelling a name and saying "D as in Denmark," phonetic features of the sound carrying 'D' may have to be held in memory until the word 'Denmark' is recognized, which makes it possible to recognize 'D' as opposed to 'T'.

Other short-term memory operations are associated with the time it takes to produce speech (selecting the appropriate words and syntactic structures and implementing the speech plan), during which interval the idea or information to be worded has to be maintained in memory.

Still others may be due to individual characteristics of a given speaker and/or his/her speech: if the speech is unclear because of its logic, information density, unusual linguistic structure or speaker's accent, the interpreter may wish to wait for a short while before reformulating it (in simultaneous) or taking notes (in consecutive) so as to have more time and a larger context to deal with comprehension and reformulation difficulties.

Language-specific factors may also require short-term memory operations. Inversions in determination sequences are one example, for instance in "System and application strategy" (from a Data Processing conference). The sequence was translated into French as "*stratégie en matière de systèmes et d'applications*." The sound and/or meaning of "System" and "application" had to be stored in short-term memory until after the interpreter heard the English word "strategy" and had said "*stratégies en matière de*".

Short-term memory operations fall under the category of non-automatic operations because they include the storage of information for later use (see Richard 1980, cited earlier). Furthermore, stored information changes both from one speech to another and during speeches as they unfold, and both stored information quantities and storage duration can vary from moment to moment, so that there is little chance for repetition of identical operations with sufficient frequency to allow automation of the processes.

3. Working memory

As explained in Chapter 1 and elsewhere in this book, in order to optimize the basic concepts and models for training purposes, they are kept simple and I have tried to avoid theoretical components to which students cannot relate in their daily interpreting experience. However, over time, the concept of working memory from cognitive psychology has come up again and again in the analyses of interpreting found in the literature. It also helps gain better understanding of the Effort Models presented later in this chapter and is useful in the analysis of some language issues (see Chapter 9). It is therefore briefly mentioned here.

Psychologists traditionally make a distinction between long-term memory (what we refer to as 'memory' in everyday life) and short-term memory, which is the ability to keep information and process it over a short period. A third type of memory, called sensory memory, as well as the interaction between the three, will be introduced in Chapter 9. Short-term memory has been investigated by many cognitive psychologists

over the past decades. It is considered an important determinant of cognitive operations and is now sometimes referred to as 'working memory' (though 'short-term memory' is a fairly generic concept while 'working memory' is more specific). In 1974, Baddeley and Hitch developed a model of working memory with a specific structure and operational components, including a 'Central Executive', a modality-free cognitive mechanism which coordinates the operation of the other entities in the model, namely a 'phonological loop' which holds information in phonological form and a 'visuo-spatial sketch pad' specialized in spatial and/or visual information coding. Baddeley and Hitch's model is described in most introductory books on cognitive psychology – see for example Eysenck and Keane 1990. Further research in the field naturally led to tests of its ability to explain and predict cognitive phenomena and then to other models with further ideas about its components, including specialized verbal working memory (Caplan & Waters 1998), about its operation, about is relationship to long-term memory. According to Miyake and Shah (1999), several ideas and theories about the components and operation of working memory compete in cognitive psychology circles (for a recent review, see Timarová 2008), and some authors even doubt the usefulness of the concepts of working memory as a separate entity, but a consensus can be found with respect to the following points:

1. Working memory is a set of mechanisms or processes involved in the control, regulation and active maintenance of task-relevant information in the service of complex cognition; it operates primarily on currently 'activated' information from long-term memory.
2. Working memory requires processing capacity
3. Working memory has a small storage capacity

As explained in more detail in Chapter 9, working memory is necessarily part of the language-comprehension process and of the speech-production process. It is obviously part of the Memory Effort and perhaps conceptually very close to it, but the Memory Effort is explained here in such a way as to be intuitively recognizable by students and professional interpreters as a step in the interpreting process which involves memory and memory operations, not as a conceptual entity from cognitive psychology. Readers may consider the Memory Effort as corresponding to working memory if they wish, but for reasons explained in Section 13, I prefer to talk about short-term memory and about the Memory Effort when referring to the mechanisms of interpreting in general and to invoke working memory only in more technical considerations.

4. An Effort Model of simultaneous interpreting

4.1 A first view of the model

Using these definitions, simultaneous interpreting (SI) can be modelled as a process consisting of the three core Efforts described above, namely the Listening and Analysis