Using on-line processing methods in language acquisition research

Theodoros Marinis
University of Reading

Running Head: On-line sentence processing methods

Address for correspondence:
Theodoros Marinis
School of Psychology & Clinical Language Sciences
University of Reading
Reading RG6 6AL
UK
Tel. +44-118-378 7465
e-mail: t.marinis@reading.ac.uk
1 Introduction

This chapter will present three widely used on-line sentence processing tasks (word monitoring, self-paced reading/listening, cross-modal priming) and one new task (self-paced listening and picture verification task) that can be used with adults and children with typical and atypical language development. But first, I would like to introduce some general characteristics of on-line sentence processing tasks and their advantages in comparison to off-line tasks.

To date, most studies investigating language comprehension have used off-line experiments to address how children and adults with typical and atypical language development comprehend words or sentences (for an overview of off-line comprehension tasks, and a detailed discussion of their advantages and disadvantages see Schmitt and Miller, this volume, and for an overview of the eyetracking technique, see Sedivy, this volume). A typical example of an off-line sentence comprehension task is the sentence-picture matching task, e.g. TROG 2 (Bishop, 2003). In this task, participants are presented with a series of pictures (usually two or four), they listen to one sentence and they have to decide which picture goes with the sentence. This can be a very powerful task if the pictures and sentences are well-designed. However, off-line tasks, such as the sentence-picture matching task, have several limitations that can make the interpretation of the data difficult. One important limitation is that off-line sentence comprehension tasks measure how participants interpret a sentence after they have heard the complete sentence. This poses a demand on the working memory of the participants because they have to process the sentence on-line as they hear it, they have to keep it in working memory and then choose the picture that matches the sentence. Participants with a relatively high working memory may perform better than participants with a relatively low working memory, so in such a task, working memory may act as a confounding factor. A second important limitation relates to the type of knowledge that they tap into. In an off-line task, participants can take time to think about the meaning of the sentence before they make a
decision and point to a picture. They make a conscious and controlled decision about the meaning of the sentence by using their explicit knowledge about language and their metalinguistic abilities. This may overestimate or underestimate their language abilities depending on whether or not they have good metalinguistic abilities.

On-line comprehension tasks, on the other hand, are relatively immune to metalinguistic abilities because they measure the participants’ unconscious and automatic response to language stimuli, although this depends on the specific task used and whether or not participants are asked to judge the acceptability of sentences. Participants do not have time to think about the meaning of the sentence and do not use their explicit knowledge about language. In addition, on-line comprehension tasks measure the participants’ responses as they listen to the sentences unfold and not at the end of the sentences. Therefore, they pose fewer demands on working memory. On-line methods also have disadvantages. They require a substantially longer time to prepare than off-line experiments because the experimental items have to be controlled for many factors; they require more experimental items, and therefore, on-line experiments tend to last longer than off-line experiments. Data analyses are more complex than in off-line tasks, and they require equipment and special software to program and run them. These are some of the general advantages and disadvantages of on-line vs. off-line experiments. Further advantages and disadvantages of on-line tasks will be discussed at the end of this chapter.

2 Word-monitoring

2.1 Rationale

In the word-monitoring task, at the beginning of each trial, participants are presented with a word aurally, e.g. *cakes*, or they see a picture representing a word, e.g. a picture with cakes. Then they
listen to sentences that may or may not contain this word, as shown in (1) below, and they have to press a button as fast as possible if they hear the word they saw or heard earlier.

(1) John’s mother is a great baker. Most days she bakes *cakes* for him and the whole family.

The computer records the reaction time (RT) of the button press for the word *cakes*, and this is the dependent variable. The relevant comparison is between the RTs to a word, as in (1) above, and the RTs to the same word in a different sentential context. This can involve ungrammaticality, as in (2) below, semantic anomaly, as in (3) below, or any other type of anomaly.

(2) John’s mother is a great baker. Most days she *bake* cakes ….

(3) John’s mother is a great baker. Most days she *drinks* cakes ….

The word to be monitored is not the word that causes the ungrammaticality/anomaly, but the word *after* the ungrammaticality. The basic assumption here is that we unconsciously slow down whenever we are faced with an ungrammaticality/anomaly. Therefore, if participants process the ungrammaticality, they should show longer RTs for the word following it, e.g. the word *cakes* in (2) compared to the same word in a grammatical sentence, e.g. the word *cakes* in (1). A crucial point is that participants do not monitor the word causing the ungrammaticality, therefore, the RTs reflect an unconscious rather than a controlled process. Off-line comprehension questions can be used to ensure that participants pay attention to and comprehend the sentences.
2.2 Linguistic variables

This task has been used to investigate the sensitivity to an ungrammaticality in terms of the omission of grammatical morphemes (Montgomery & Leonard, 1998; Montgomery & Leonard, 2006), as shown in (1) vs. (2) above, and also the sensitivity to syntactic, semantic, and world-knowledge information (Montgomery, Scudder & Moore, 1990; Tyler & Marslen-Wilson, 1981; Komisarjevsky Tyler, 1992). The word-monitoring task can be used to test any type of ungrammaticality or anomaly that becomes obvious immediately prior to the word that is being monitored. This task is less sensitive when there is a distance between the ungrammaticality and the word that is being monitored.

2.3 Subjects

This method can be used by adults and children with normal or corrected to normal sight/hearing. To date, it has been used successfully with adults, children from the age of 5 onwards (Tyler & Marslen-Wilson, 1981), language impaired adults (Komisarjevsky Tyler, 1992), and children with Specific Language Impairment from the age of 6 onwards (Montgomery & Leonard, 1998; Montgomery & Leonard, 2006). It could be used with any population that is able to understand the task of monitoring a word and pressing a button. It is therefore appropriate for monolingual and bilingual adults and children from about the age of 5. However, given that participants have to keep a word in their memory and press the button when they hear it, participants with low working memory or problems with lexical access may show a lower accuracy rate in pressing the button and/or slower reaction times compared to adults with typical language abilities. This difference could be due to working memory limitations and problems with lexical access and not because they do not process the ungrammaticality or anomaly in the sentence. This is also possible for bilinguals with a relatively low proficiency level if the sentences include unfamiliar vocabulary or if they have
slow lexical access and/or difficulties to comprehend spoken language due to their level of proficiency. Also, from the experience of using this task in our lab, some 5 year-old children have difficulties understanding and performing the task. This may relate to the fact that this is a dual task – monitoring a word and comprehending a sentence at the same time. This method is clearly inappropriate for participants with a motor impairment because the data come from pressing a button.

2.4 Description of procedure

The word monitoring task requires three types of sentences: 1) experimental sentences, 2) filler sentences, and 3) catch trials. Experimental sentences include the grammatical and ungrammatical sentences that are testing the research question(s) of the study. Filler sentences are used so that the participants are made unaware of the purpose of the task and to avoid participants becoming familiarized with the types of sentences and ungrammaticality used. Studies with adults include an equal number of filler and experimental sentences, and sometimes more filler than experimental ones. Studies with children include usually a smaller number of filler sentences in order to reduce the length of the experimental session and to avoid fatigue and loss of attention. Catch trials are trials that do not include the word to be monitored, and thus, participants should not press the button. These are required to make sure that the participants do not press the button mechanically and to keep them alert and attending the task. To ensure that participants not only monitor the word, but also pay attention and comprehend the sentences, comprehension questions should be used for a proportion of the experimental sentences, fillers, and catch trials.

Usually, trials consist of two sentences, a leading sentence that creates a context and the test sentence. Lead-in sentences facilitate the interpretation of the test sentences (Marslen-Wilson & Tyler 1981). Therefore, although they are not necessary, they are recommended especially in
experiments with children and impaired populations. Test sentences should be constructed very carefully, taking into account several factors, such as the length and structure of the sentence, the position of the word to be monitored, the structural relation between the critical word and the word that is being monitored, the properties of the critical word and the word that is being monitored, such as syntactic category, length in syllables, frequency, age of acquisition of the word, imageability, and other properties that relate to the research questions and design of the particular study. The monitored word is usually a noun, but it can also be a verb or adjective (see Komisarjevsky Tyler, 1992). It can be identical to the word presented at the beginning, a member of a semantic category, a rhyme, or of similar orthographic shape (Marslen-Wilson & Tyler, 1980).

Test sentences should have similar length and structure to avoid differences in length and complexity of the sentences to interfere with the design of the study. For example, all test sentences should have the same number and type of clauses, arguments, adjuncts. The word that is being monitored should not be at the same position in each sentence. This is to avoid participants developing an expectation that they have to press the button at the beginning, middle, or towards the end of the sentence. The position can be defined based on the number of words or syllables. The structural relation between the critical word and the word that is being monitored and the distance between the two should be kept constant, otherwise a confounding factor will be introduced into the design. Similarly, the properties of the critical word and the word that is being monitored should also be controlled for. Differences in the syntactic category of the word, length in syllables, age of acquisition of the word, and imageability may cause longer or shorter RTs.

Finally, if the words that are being monitored are presented in pictures prior to the sentences, these pictures should be controlled for several factors, such as picture complexity, speed of naming, and a pretest naming task should be performed to ensure that the pictures elicit the words that will be monitored in the word monitoring task.
In the WMT, the monitored word cannot appear more than once in a single session due to repetition-priming effects. Therefore, the monitored words need to be counterbalanced across different sessions and across conditions (Tyler, 1992).

The equipment required for the word monitoring task include a desktop or laptop computer, software for presentation of the stimuli and recording of RTs, and a button box for collecting RT data. If the word to be monitored is presented in a picture, an LCD or CRT monitor will be required. Button boxes developed for recording RTs in milliseconds are preferred rather than using a mouse or the space bar on the keyboard because they are more accurate and consistent across trials. Two sets of headphones are required for the presentation of the auditory stimuli, one for the participant and one for the experimenter.

2.5 Analysis and outcomes

The data generated in the word monitoring task are of two types: 1) accuracy in monitoring the word, and 2) RTs in milliseconds on how fast participants pressed the button. The first step in the data analysis is to calculate the accuracy data in percentage per condition. Trials in which the participants did not detect the word or did not press the button should be excluded from further analyses. Data should also be excluded in trials in which participants pressed the button to a non-target word. In relatively homogeneous populations, such as non-impaired adults and typically developing children, participants whose accuracy is more than 2 standard deviations below the mean of the group could be considered as outliers and excluded from further analyses in order to reduce noise in the data. This is more controversial in clinical populations and second language learners because of the high degree of heterogeneity in the groups.

After the analysis of accuracy, it is crucial to screen the RT data for extreme values and outliers (Ratcliff, 1993). Both are defined using a cut-off point, which differs for different groups of
participants depending on their speed in pressing the button. Children often have longer RTs than adults; children with language impairment have longer RTs than typically developing children of the same age. The variance in children’s data is also often larger than the variance in adult’s data; the variance in the data of impaired populations is larger than in non-impaired ones. Extreme values can be best defined on the basis of the RTs in the particular experiment and experimental group. Outliers are usually defined as data points above and below 2 standard deviations. Outliers should be calculated for each condition separately per participant and per item. Extreme values and outliers can be either excluded from the analysis or they can be replaced with the mean value in each condition per participant or per item. The latter is a more conservative way of trimming the data. In relatively homogeneous populations, participants whose overall RTs are greater than 2 standard deviations from the mean of the group could also be considered as outliers and excluded from further analyses to decrease the amount of noise in the data.

RT data are often not normally distributed. This is because there is a limit as to how fast participants can press the button, so the data are positively skewed. A log transformation can be used to transform the data prior to analyzing them using parametric tests. Depending on how many factors have been used in the task, the data can be analyzed using t-tests (for 1 factor with 2 levels, for example grammatical vs. ungrammatical), or repeated measures ANOVAs (for more complex designs). Either way, RT experiments require analyses per participants and per items. If the analyses per items shows different results from the analysis per participants, this is often because there is more variance in the items than in the participants or the other way around.

Children’s RTs become faster as they grow older, and adults’ RTs become slower as they age. Therefore, RT studies with a large age range of participants will show large standard deviations and small effects may not reach significance level. A way around this problem is to have an independent
task that measures mean RTs per participant. This can then be used as a co-variable to partial out individual differences in RTs, see Montgomery & Leonard (2006).

The outcomes from the accuracy analysis show how good participants are in monitoring words, which relates to attention and lexical access. RT data show whether or not participants notice an ungrammaticality or anomaly in the sentence.

3 Self-paced reading/listening

3.1 Rationale

In the self-paced reading and listening tasks, participants read/listen to sentences in a word-by-word or phrase-by-phrase fashion by pressing a button. Participants have control over the rate of presentation of each sentence; therefore, the task is self-paced. Each button press is recorded, and thus, provides insight into how fast participants process each word or phrase. Longer RTs at particular positions in a sentence are thought to reflect processing difficulties, which could relate to the ungrammaticality of the sentence, violation of an expectation, or a reanalysis process (Just, Carpenter & Wooley, 1982).

3.2 Linguistic variables

This task has been used widely in the past to investigate several phenomena, such as the processing of temporarily ambiguous sentences (e.g., Ferreira, Anes & Horine, 1996a; Ferreira et al., 1996b; Garnsey, Pearlmutter, Myers, & Lotocky, 1997; Juffs, 1998; Felser et al., 2003a; Felser et al., 2003b; Fernández, 2003; Papadopoulou & Clahsen, 2003; Jackson, 2008), filler-gap dependencies (e.g., Stowe, Tanenhaus, & Carlson, 1991; Pickering & Traxler, 2003; Aoshima, Phillips, & Weinberg, 2004; Marinis et al., 2005), and pronoun resolution (e.g., Carminati, 2005; Wolf, Gibson
& Desmet, 2004; Stewart, Holler & Kidd, 2007). These studies have revealed that parsing is guided or constrained by different types of information, such as phrase-structure information, lexical-semantic information, prosodic information, and contextual information (Gibson & Pearlmutter, 1998; Gibson et al., 1996).

3.3 Subjects

The self-paced reading/listening tasks can be used by both adults and children with normal or corrected to normal sight/hearing. The self-paced reading task is suitable for populations that have well-developed literacy skills because the inferences we make about processing are mediated through reading. The bulk of studies using this methodology has been with adult native speakers of various languages. Recently, this task has also been used with adult second language learners (Felser et al., 2003b; Juffs & Harrington, 1995; Juffs & Harrington, 1996; Marinis et al., 2005; Papadopoulou & Clahsen, 2003). Some recent studies have also used the self-paced reading methodology with 8 to 12 year-old children (Booth, MacWhinney & Harasaki, 2000; Papadopoulou & Tsimpli, 2005; Traxler, 2002). The self-paced listening task can be used by both children and adults (Booth et al., 2000; Felser, et al., 2003a; Ferreira et al., 1996b; Kidd & Bavin, 2007).

3.4 Description of procedure

The self-paced reading task can have three types of presentation: 1) cumulative presentation, 2) linear non-cumulative presentation, and 3) centre non-cumulative presentation. In the cumulative and linear non-cumulative presentation, participants first see dashes on the computer screen that correspond to the letters of the words of the sentence. When the sentence starts, they see the first word/phrase on the screen. When they press the button, the first word/phrase remains on the screen and the second one appears. As the sentence progresses, previous words/phrases remain on the
computer screen and participants can go back and read previous words/phrases. An example of this presentation sequence is provided in (4) below for the sentence ‘The doctor examined / the nurse of the pupils / who / was / feeling very tired’ (Felser et al., 2003b).

(4) Cumulative self-paced reading presentation

--- ------- --- ----- --- ----- --- ------- --- ------ --- ----- --- ------ --- ------.
The doctor examined --- ----- --- ---- --- ------- --- ------ --- ------ --- ------.
The doctor examined the nurse of the pupils --- -------- --- ------ --- ------ --- ------.
The doctor examined the nurse of the pupils who --- ------- --- ------.
The doctor examined the nurse of the pupils who was -------- --- ------.
The doctor examined the nurse of the pupils who was feeling very tired.

In the linear non-cumulative presentation, words/phrases disappear each time a new word/phrase appears on the computer screen, and so participants cannot read previous words/phrases again. An example of this presentation sequence for the same sentence is given in (5) below.

(5) Linear non-cumulative self-paced reading presentation

--- ------- --- ----- --- ----- --- ------- --- ------ --- ----- --- ------ --- ------.

The doctor examined --- ----- --- ---- --- ------- --- ------ --- ------ --- ------.

--- ------- ------- the nurse of the pupils --- -------- --- ------ --- ------.

--- ------- ------- --- ----- --- ---- --- ------- --- ------ --- ------ --- ------.

--- ------- ------- --- ----- --- ---- --- ------- --- ------ --- ------ --- ------.

--- ------- ------- --- ----- --- ---- --- ------- --- ------ --- ------ --- ------.

--- ------- ------- --- ----- --- ---- --- ------- --- ------ --- ------ --- ------.

--- ------- ------- --- ----- --- ---- --- ------- --- ------ --- ------ --- ------.

--- ------- ------- --- ----- --- ---- --- ------- --- ------ --- ------ --- ------.

--- ------- ------- --- ----- --- ---- --- ------- --- ------ --- ------ --- ------.

--- ------- ------- --- ----- --- ---- --- ------- --- ------ --- ------ --- ------.

--- ------- ------- --- ----- --- ---- --- ------- --- ------ --- ------ --- ------.

--- ------- ------- --- ----- --- ---- --- ------- --- ------ --- ------ --- ------.

--- ------- ------- --- ----- --- ---- --- ------- --- ------ --- ------ --- ------.

--- ------- ------- --- ----- --- ---- --- ------- --- ------ --- ------ --- ------.

--- ------- ------- --- ----- --- ---- --- ------- --- ------ --- ------ --- ------.

--- ------- ------- --- ----- --- ---- --- ------- --- ------ --- ------ --- ------.

--- ------- ------- --- ----- --- ---- --- ------- --- ------ --- ------ --- ------.

--- ------- ------- --- ----- --- ---- --- ------- --- ------ --- ------ --- ------.

--- ------- ------- --- ----- --- ---- --- ------- --- ------ --- ------ --- ------.

--- ------- ------- --- ----- --- ---- --- ------- --- ------ --- ------ --- ------.

--- ------- ------- --- ----- --- ---- --- ------- --- ------ --- ------ --- ------.

--- ------- ------- --- ----- --- ---- --- ------- --- ------ --- ------ --- ------.

--- ------- ------- --- ----- --- ---- --- ------- --- ------ --- ------ --- ------.
The centre non-cumulative presentation is similar to the linear non-cumulative presentation, but in this variant of the task the words/phrases appear in the centre of the computer screen, as shown in (6) below.

(6) Centre non-cumulative self-paced reading presentation

The doctor examined

the nurse of the pupils

who

was

feeling very tired.

Non-cumulative presentation provides a more accurate picture of how participants process sentences on-line compared to the cumulative presentation because in the non-cumulative presentation participants cannot go back and read parts of the sentence again. However, the cumulative presentation is more similar to the way we read sentences in real life. One disadvantage of the linear non-cumulative presentation is that participants can see how long the sentence is based on the dashes; they can see how close each word they read is to the end of the sentence. Knowledge of the length of a sentence and how close a word is to the end of the sentence can cause the development of expectations and predictions about the incoming words. This is impossible in the centre non-cumulative presentation because participants can see only one word/phrase at a time at the centre of the screen and they do not have any clues about the length of the sentence. This presentation type is also more similar to the way we process language when we listen to sentences.

In the self-paced listening task, sentences can be presented word-by-word or phrase-by-phrase. There is only one mode of presentation, i.e., participants listen to each word or phrase one at a time.
without knowing how long the sentence is and without being able to go back and listen again to any of the words/phrases they have already heard. This is why the mode of presentation in the self-paced listening task is similar to the centre non-cumulative presentation of the self-paced reading task.

Studies using the self-paced reading/listening procedure have often included comprehension questions at the end of either each sentence or at the end of a proportion of the sentences. This is in order to keep participants focused on the task of reading and comprehending the sentences instead of pressing the button mechanically. This can provide additional accuracy data for the comprehension of the experimental sentences and fillers. Based on the comprehension accuracy data it is possible to analyze separately the RTs of correctly and incorrectly comprehended sentences (Juffs & Harrington, 1996).

As mentioned in the word monitoring task, filler sentences are used to disguise the purpose of the task from participants and to avoid developing expectations and strategies. The number of fillers can vary with more fillers in studies with adults than with children for practical reasons. The material has to be constructed very carefully to avoid confounding factors interacting with the design of the study.

In the self-paced reading and listening tasks, we measure RTs for each word/phrase of the sentence. Therefore, the sentences should have similar length and structure to avoid differences in length and complexity that may act as confounding factors of the design. There are usually one or two critical segments in each sentence that provide the crucial information for our research question. These have to have exactly the same words or form minimal pairs, depending on the design of the study. Using different words will introduce confounding factors in the design. For example, in the Marinis et al. (2005) self-paced reading study on the processing of intermediate
traces in sentences with successive cyclic movement, the critical parts of the sentence were Segments 3 and 5, as shown in (7) below.

<table>
<thead>
<tr>
<th>Seg1</th>
<th>Seg2</th>
<th>Seg3</th>
<th>Seg4</th>
<th>Seg5</th>
<th>Seg6</th>
</tr>
</thead>
<tbody>
<tr>
<td>(7a)</td>
<td>The manager who/ the secretary claimed/ $t_2$ that/ the new salesman/ had pleased $t_1$/ will .</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(7b)</td>
<td>The manager who/ the secretary’s claim/ about/ the new salesman/ had pleased $t_1$/ will .</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(7c)</td>
<td>The manager thought/ the secretary claimed/ that/ the new salesman/ had pleased/ the ..</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(7d)</td>
<td>The manager thought/ the secretary’s claim/ about/ the new salesman/ had pleased/ the ..</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

According to generative linguistic theory, who in sentences (7a) and (7b) moves from the object position of the verb pleased in the subordinate clause to the main clause leaving a trace ($t_1$) or gap behind. Sentence (7a) involves successive cyclic movement, so who moves first to the beginning of the subordinate clause leaving an intermediate trace ($t_2$) or gap behind and then it moves to the main clause. Movement in (7a) and (7b) creates a syntactic dependency between who (the filler), the intermediate trace/gap ($t_2$), and the trace/gap after the verb ($t_1$). According to the Trace Reactivation Hypothesis (Swinney et al., 1988), when we encounter a filler, we store it in short-term memory and we try to integrate it as soon as possible into the sentence. When the parser identifies a gap, i.e. a potential position for integration, it retrieves the filler from short-term memory and sets up a filler-gap dependency by reconstructing the grammatical and semantic features of the filler at the position of the gap. Based on this hypothesis, who in sentence (7) will be stored in short-term memory. At the position of each trace/gap its grammatical and semantic features will be reconstructed, a process that requires processing resources and is reflected in longer RTs. Thus, in example (7) above, RTs
at Segment 5 (*had pleased*) in (7a) and (7b), which involve a trace/gap, are expected to be longer than in (7c) and (7d), which do not involve a trace. Similarly, RTs at Segment 3 (*that*) in (7a), which involves an intermediate trace, are expected to be longer than in (7c), which does not involve a trace.

In the self-paced listening task, prosody can provide cues for the structure of the sentence (Ferreira et al., 1996a). If the study does not want to test how participants make use of prosodic cues, then prosodic cues that may affect the critical segments have to be eliminated, otherwise they will be confounding factors in the design. Eliminating of prosodic cues can be done through splicing, i.e., replacing words that include prosodic cues (Felser et al., 2003a), by changing the prosody using computer software, through recording the words as lists, or through recording the words in flat intonation. However, it should be stressed that word-by-word or phrase-by-phrase segmentation and eliminating of prosodic cues is very likely to cause unnatural speech, which may affect the research questions and hypotheses tested.

The equipment required is similar to the word monitoring task, but there is a crucial difference for the self-paced reading task. Given that in this task we measure RTs upon reading words/phrases on the computer screen, it is important to use a computer screen with a fast refresh rate that can be controlled by the software. This can only be achieved by using CRT monitors. Laptop monitors and LCD monitors are not recommended for self-paced reading experiments, especially in experiments using short duration stimulus displays.

### 3.5 Analysis and outcomes

The procedure for the analysis of accuracy data and screening the data for extreme values and outliers is the same as in the word monitoring task. Trails from inaccurate responses are usually
excluded from further analyses because it is unclear whether or not the participants comprehended the sentences in those trials.

There are two possible ways to analyze the RT data from the self-paced reading/listening task: 1) analyze the raw RTs as recorded by the software, or 2) analyze residual RTs. residual RTs take into consideration individual variation between participants and are calculated on the basis of the length of each word/phrase and the raw RTs for each participant (for more details, see Ferreira & Clifton, 1986; Trueswell, Tanenhaus & Garnsey, 1994). The advantage of residual RTs is that it cancels out individual differences of speed between the participants. Residual RTs should be calculated if in the critical segment there are differences in the length of words in the self-paced reading task or in the duration of the words in the self-paced listening task.

The outcomes from the accuracy analysis show how well participants comprehend sentences. This provides information for off-line comprehension. RT data reveal at which words in the sentence participants encountered processing difficulties, which are reflected in longer RTs.

4 Cross-modal priming

4.1 Rationale

The cross-modal priming paradigm is a dual task, i.e. participants have to perform two different tasks during the same procedure which also involves two modalities (auditory and visual). It has been used in the past to investigate lexical processing, for example the processing of homophones (Grainger, Diependaele, Spinelli, Ferrand, & Farioli, 2003) and syntactic processing, for example wh-dependencies (e.g., Love & Swinney, 1996; Love & Swinney, 2007; Roberts, Marinis, Felser & Clahsen, 2007), and pronoun resolution (e.g., Nicol & Swinney, 1989; McKee, Nicol & McDaniel, 1993). In this task, participants listen to a sentence and at a relevant point in the sentence they see a word (cross-modal lexical priming) or a picture (cross-modal picture priming) on the computer.
screen. Upon seeing the word or picture, they have to name it, make a lexical decision, or picture classification task (e.g. an animacy task) by pressing a button as fast as possible. The button press is recorded and provides the data of the cross-modal priming task. For example, sentence (8) below consists of a main clause and a relative clause. The object in the main clause (peacock) is the indirect object in the relative clause, so the sentence involves movement of the indirect object peacock from the relative clause to the main clause leaving a trace (t₁) or gap behind, which creates a syntactic dependency between the word peacock (the filler) and the trace or gap.

(8) John saw the peacock, [to which the penguin gave the nice birthday present t₁ in the garden]

As mentioned in section 2.4 above, fillers are thought to be stored temporarily in short-term memory; when the parser identifies a gap, i.e. a potential position for integration, it retrieves the filler from short-term memory and sets up a filler-gap dependency by reconstructing the grammatical and semantic features of the filler at the position of the gap. Based on this hypothesis, peacock in sentence (8) will be stored in short-term memory. At the position of the trace/gap its grammatical and semantic features will be reconstructed. Thus, if the word peacock or the picture of a peacock is presented at the trace/gap and participants have to name it, to make a lexical decision task or a picture classification task, RTs will be shorter compared to presentation of the same picture in a different position in the sentence and compared to the RTs to an unrelated picture. So, the rationale in the cross-modal priming experiment is that RTs at a syntactic relevant position will be shorter because of the reactivation of grammatical and semantic features.

(Roberts et al., 2007)
4.2 Linguistic variables

The cross-modal priming paradigm can be used for the investigation of any structures involving a dependency between two constituents. In the past it has been used extensively to investigate the processing of filler-gap dependencies involving wh-movement (Love & Swinney, 1996; Love & Swinney, 2007; Marinis & van der Lely, 2007; Nicol, 1993; Roberts et al., 2007), object scrambling (Clahsen & Featherston, 1999; Nakano, Felser & Clahsen, 2002), and referential dependencies, e.g., reference of pronouns and reflexives (McKee et al., 1993; Nicol & Swinney, 1989).

4.3 Subjects

The cross-modal priming paradigm has been used with non-impaired adults (Love & Swinney, 1996; Nakano et al., 2002), adults with aphasia (Balogh et al., 1998; Blumstein et al., 1998; Swinney et al., 1996; Zurif et al., 1993), adult second language learners (Roberts et al., 2007), typically developing children from the age of 4 (Love & Swinney, 2007; McKee et al., 1993; Roberts et al., 2007), and children with Specific Language Impairment (Marinis & van der Lely, 2007). However, the study by Roberts et al. (2007) has shown that children and adults with a relatively low working memory do not show antecedent reactivation. Therefore, it is wise to use a measure of working memory together with the cross-modal priming task in order to be able to control for working memory. Also, the cross-modal priming task is a demanding task for young children because it is a dual task that requires attention in two separate tasks. Therefore, young children require a lot of practice to familiarize themselves with the task in order to be able to perform it appropriately. A good practice is to teach first the naming, lexical decision or picture classification part of the task. A second step would be to familiarize the participants with performing the naming, lexical decision, or picture classification while listening to a sentence.
However, from the experience in our lab, despite practice, many 4 and 5-year old children find it very difficult to perform this task.

### 4.4 Description of procedure

In the cross-modal priming paradigm, while participants listen to a sentence, they see a word or a picture on the computer screen, which is referred to as the target, and they have to do a secondary task related to the target. This can be naming, lexical decision, or picture classification, e.g. participants have to press one of two buttons to decide whether the word on the computer screen is a real word or a non-word, or to classify whether a picture shows a living or a non-living thing. Half of the targets are related to a word in the sentence (related targets) and the other half are unrelated (unrelated targets).

The related targets can be strong semantic associates or identical repetitions of the word in the sentence. Identical repetitions produce a stronger priming effect than semantic associates. Using semantic associates also involves an additional process to establish a semantic association between the word/picture and the antecedent (Clahsen & Featherston, 1999). A disadvantage of using identical targets is that participants may become conscious that the word/pictures were mentioned earlier in the sentence. A large number of fillers can be used to avoid this (Clahsen, 2008).

The visual stimuli are presented at a syntactically relevant position, and also at a control position that usually precedes the syntactically relevant position. For example, in the Marinis & van der Lely study the syntactically relevant position is the trace/gap, indicated with [3] in example (9) below. The control position is earlier in the sentence, at the offset of an adjective, indicated with [2]. This study tested also possible effects at the offset of the verb, indicated with [1]. This study used as visual stimuli pictures of identical and unrelated targets. In the example below, the identical target was the picture of a rabbit and the unrelated target was the picture of a ladder.
In the cross-modal priming paradigm the prediction is that the RTs at the syntactically relevant position will be shorter for semantic associates or identical targets compared to unrelated targets because the grammatical and semantic features of the filler will be reconstructed at the gap. This by itself does not provide evidence for a syntactic dependency. Shorter RTs for semantic associates and identical targets compared to unrelated targets could be caused because the semantic associates and identical targets are related to a word that was mentioned earlier in the sentence. Evidence for a syntactic dependency can be provided if this difference in RTs is not attested at the control position.

Given that the relevant comparison in this task is between the RTs for two different words/pictures, for example in (9) above between RTs for rabbit vs. ladder, the words that are compared must be matched on several factors that may potentially affect RTs, such as frequency, age of acquisition of the word, length in terms of number of letters, syllables, and neighborhood density. If visual stimuli consist of pictures, these should be of similar complexity and pre-testing of the material through a naming task should ensure that they correspond to the related word in the sentence. The picture pairs should also be matched on imageability and pre-testing should ensure that the pictures in each pair have similar speed of naming.

Similarly to the self-paced reading task, the cross-modal priming task requires the use of a CRT monitor.
4.5 Analysis and outcomes

As in the word monitoring task, the first part of the analysis involves analyzing accuracy data. In the cross-modal priming task, these consist of accuracy data for the secondary task, i.e. naming, lexical decision, or picture classification. These data show how accurate participants were in naming, lexical decision, or picture classification, so they do not provide information about sentence comprehension and the processing of syntactic dependencies. Inaccurate responses should be excluded from the RT analyses.

RT data should be analyzed in the same way as in the word monitoring task and self-paced reading/listening task. The outcome of the RT analyses can reveal how participants process syntactic dependencies.

5 Self-paced listening with picture verification

5.1 Rationale

This is a new child-friendly task that we developed recently (Marinis, 2007; Marinis, 2008; Marshall, Marinis & van der Lely, 2007) and consists of two tasks combined, i.e., self-paced listening and picture verification. Participants see a picture on the computer screen and then they listen to a sentence in a word-by-word or phrase-by-phrase fashion, similarly to a self-paced listening task. Sometimes, the sentence matches to the picture, and sometimes it does not. For example, the study by Marinis (2007) involved active and passive sentences in English. Participants saw a picture with an action, e.g., a zebra kissing a camel, and they heard the following sentences:

(10a) I think / that / the zebra / was kissing / the camel / at the zoo / last Monday.

(10b) I think / that / the zebra / was kissed / by the camel / at the zoo / last Monday.
This is a minimal pair. Until the phrases ‘was kissing’ and ‘was kissed’ it is ambiguous as to whether there is a match or mismatch between the picture and the sentence. The disambiguating information is encoded in the inflectional suffixes –ing and –ed. The rationale is that if participants are processing the inflectional suffixes and assign thematic roles rapidly before the end of the sentence, RTs on the mismatch condition will be longer than in the matching condition.

5.2 Linguistic variables

To date, this paradigm has been used to investigate the processing of English active and passive sentences (Marinis, 2007; Marshall et al., 2007), pronoun resolution in English reflexive and non-reflexive pronouns (Marinis, 2008), processing of active and non-active morphology in Greek (Papangeli, in progress), and pronoun resolution in Greek (Papadopoulou et al., 2007). This task can be used for any phenomenon for which minimal pairs as the ones above can be constructed to test rapid integration of different types of information.

5.3 Subjects

This task is very easy to explain and perform. Therefore, it can be used with children, adults, and impaired populations. The youngest children in the studies above were 5 years of age, but in principle it can also be used with younger children.

5.4 Description of procedure

As mentioned above, participants first see a picture on the computer screen and then listen to a sentence word-by-word or phrase-by-phrase by pressing a button. In the studies above, the picture was presented prior to the sentence in order for children to create a mental representation of the scene and also to create expectations about the following sentence. At the end of the sentence,
participants had to indicate off-line whether there is a match or mismatch between the picture and the sentence. This keeps participants on task, and also provides off-line data on how accurate participants are in comprehending the sentences.

The pictures and sentences have to be created very carefully to avoid confounding factors. For example, in the Marinis (2007) study, the pictures consisted of actions between animate participants, and they were reversible. The direction of the action was half of the time from right to left and for the other half of the time from left to right. The critical segment in the sentences was a minimal pair, controlled for duration. Verbs and nouns were controlled for frequency, age of acquisition, regularity of the past participle, and phonotactics. The critical segments were followed by two prepositional phrases, e.g., *at the zoo last Monday*, in order to be able to observe possible spill-over effects. A large number of fillers was included to avoid the development of strategies.

This task requires the same equipment with a self-paced listening task. The refresh rate of the monitor is not important because the task measures RTs on the auditory stimuli and not the pictures. Therefore, this task can be performed with a CRT or LCD monitor.

### 5.5 Analysis and outcomes

The first part of the analysis involves analyzing the accuracy data of the off-line matching task. This can demonstrate off-line comprehension of the sentences. Inaccurate responses should be excluded or analyzed separately. The next step involves screening the RT data for extreme values and outliers. RTs should be analyzed similarly to the self-paced reading/listening task. The outcome of the RT analyses shows whether or not participants detect the mismatch between pictures and sentences. Based on this finding, we can make inferences as to whether or not participants are sensitive to different types of information when they process sentences in real-time.
6 Advantages and disadvantages

The most important advantage of on-line compared to off-line tasks is that they are implicit tasks, i.e. participants are not asked to think about the structure of sentences and make acceptability judgments (unless the task includes a secondary task involving grammaticality judgment). Therefore, they are relatively immune to metalinguistic knowledge and response strategies. They measure the participants’ unconscious behavior, and thus, they can better reflect the underlying process involved when participants process language in real-time.

6.1 Word-monitoring task

The main advantage of the word-monitoring task is that it can measure the participants’ unconscious reaction to violations by directing their attention to a word that follows the violation and not the word the causes the violation. In addition, this task is easy to understand and perform, and therefore, it can be used with young children of different ages, as well as with adults, and with language impaired participants.

A disadvantage of the word-monitoring task is that it implicates working memory because participants have to keep a word in working memory and press a button when they hear it again. Therefore, it is wise to include an independent measure of working memory in order to be able to address whether or not the findings of the word-monitoring task are modulated by working memory capacity. However, it should be noted that recently many studies with impaired adults and children have used this method successfully (Montgomery & Leonard, 1998; 2006; Tyler, 1992). A second disadvantage of the task is that it requires the word that is being monitored to appear immediately after the violation. This is sometimes very difficult to do depending on the violation, and the word order of the language. A final disadvantage of this task is that it can test only the processing of violations, and therefore, it is limited to the phenomena that involve violations.
6.2 Self-paced reading/listening

As in the word monitoring task, the self-paced reading/listening tasks are implicit and are thought to measure the participants’ unconscious behavior. One advantage of the self-paced reading/listening over the word-monitoring task is that in the self-paced reading/listening task we measure RTs for each word/phrase, whereas in the word-monitoring task we measure RTs in only one point in the sentence. Thus, the self-paced reading/listening task provides more data points because it provides information about processing word-by-word or phrase-by-phrase throughout the duration of the sentence. This can reveal processing difficulties at any point within the sentence and can provide a better understanding on how participants process sentences incrementally.

Similarly to the word monitoring task, the self-paced reading task is easy to learn and perform. The self-paced listening task requires more practice in order for the participants to get used to pressing the button at the end of each auditory stimulus. If they press the button prior to the end of the auditory stimulus, they will not hear the whole word/phrase, and the trial may have to be excluded from the analysis.

One important disadvantage of the self-paced reading and listening task is that sentences are presented at a slower rate than in normal speech and the presentation is unnatural because it is segmented into words or phrases. This does not correspond to the way we read or listen to sentences. In the self-paced listening task, segmentation breaks the intonation pattern of the sentence. The presentation is even more unnatural if intonation is neutralized. If sentences are very long, this may cause problems for participants with working memory limitations because they may forget the beginning of the sentence. Finally, a disadvantage for children is that they may find the self-paced reading/listening task boring if they have to listen to many sentences without any visual stimuli. To avoid fatigue, visual stimuli can be presented at different intervals as short breaks.
6.3 Cross-modal priming

Similarly to the previous on-line tasks, the cross-modal priming task measures the participants’ unconscious behavior, and therefore, it is immune to metalinguistic knowledge. An advantage of the cross-modal priming task over the self-paced reading/listening task is that the sentence is presented unsegmented, and therefore, it is at normal speed. An advantage compared to the word-monitoring task and the self-paced reading/listening task is that it involves two modalities, visual and auditory, and therefore, the findings are minimally affected by form overlap.

The main disadvantage of the cross-modal priming task is that it is a dual task that involves decision-making, e.g. animate vs. inanimate. Therefore, this task is not as easy to learn and perform as the word monitoring task and the self-paced reading/listening task, and it requires higher levels of processing capacity and working memory. Populations with capacity limitations and low working memory may not be able to perform the task or may show a delayed priming effect because of the complexity of the task and slower speed of processing. Another disadvantage is that similarly to the word monitoring task, the cross-modal priming task measures RTs only at one point in the sentence, and not for each word/phrase as the self-paced reading/listening does.

6.4 Self-paced listening and picture verification task

The main advantage of this task compared to the other on-line processing tasks is that it is very easy to understand and perform, and therefore, it can be used with a wide range of populations including young children. A further advantage is that similarly to the self-paced reading/listening, it records RTs for each word/phrase, and enables us to observe how sentences are processed incrementally. In this task, participants can control how fast they will listen to the sentences. This is important for impaired populations that process sentences at a slower rate, for example children with Specific Language Impairment who can listen to the sentences of this task in speech rate that is optimal for
them. Finally, this task puts less memory demands on the participant than the previous mentioned on-line tasks because the picture remains on the computer screen throughout the presentation of the sentence.

The main disadvantage of the self-paced listening and picture verification paradigm is that the match/mismatch judgment between picture and sentence may make the participants develop expectations and as a consequence, response strategies. This can be avoided by introducing a large number of filler sentences. In addition, similarly to the self-paced listening task, word-by-word or phrase-by-phrase presentation of the sentences is unnatural because it breaks the intonation pattern of the sentence, and it is even more unnatural if the intonational cues are neutralized.

7 Dos and don’ts

- Do ensure that participants are naive to the purpose of the task as otherwise their response may be controlled and conscious;
- In ideal circumstances, the experimenter should not be in the same booth as the participant because this may affect the participants’ performance. This may not be possible when testing children and impaired populations;
- Ideally, the length of an on-line experiment should not exceed 15-20 minutes to avoid fatigue that may affect reaction times. If this is impossible due to the number of conditions and items, short breaks should take place;
- Do plan at least a one week gap if an on-line experiment involves several sessions to avoid repetition effects;
- Do create the stimuli very carefully controlling for several factors, such as length of sentences in words/syllables, length of words in syllables, frequency, age of acquisition, imageability of
words, complexity of sentences, intonation in auditory stimuli. This is not an exhaustive list of factors that should be controlled for. The factors that should be controlled for depend on the design of the task, the research questions, and hypotheses;

- Do collect data first from adult monolinguals in studies with young children, bilingual/multilingual participants, or impaired populations in order to find out how participants with fully developed language perform in the task.

- Do pilot the task with participants from the same population and possibly modify it in order to ensure that they can perform the task; to identify possible problematic sentences or pictures, in order to eliminate effects that are orthogonal to the research question and could obscure results;

- Do ensure that children and bilingual/multilingual populations can understand the vocabulary used by pretesting the material;

- Do take into account the proficiency level of bilingual/multilingual participants, as proficiency in the language you are testing may affect the results of the on-line task (e.g., Hopp, 2006). This may depend on the proficiency level, the task used, and the phenomenon under investigation. Therefore, an independent proficiency task is crucial. The scores of the proficiency task can be used as a co-varying factor in the data analysis or to split the participants into groups. Amount and type of exposure may also affect results from on-line tasks (Dussias, 2003; Fernández, 2003).

- Do use a background questionnaire to collect information about exposure that can be used to co-vary in the data analysis or to divide the participants into separate groups.

- Do try to avoid cognates because activation of the word in one language may affect RTs in another;
• Differences in working memory capacity may affect the results in all populations including monolingual adults, children, and L2 learners (e.g., Roberts, et. al., 2007, but see Juffs, 2004 who did not find a correlation in adult L2 learners). Therefore, it is a good practice to use a working memory task in order to co-vary the working memory scores or to split participants into groups. However, there is no consensus as to the type of working memory task that correlates with sentence comprehension (see Daneman & Carpenter, 1980; Waters & Caplan, 1996), e.g., reading/listening span, word recall, non-word repetition, digit span, word span. In our lab, we have used successfully Daneman & Carpenter’s (1980) reading span with monolingual and bilingual adults, and the listening span tasks from Gaulin & Campbell (1994) and Pickering & Gathercole (2001) with typically developing children and children with SLI. An advantage of the Pickering & Gathercole (2001) listening span task is that it has been normed for children between the ages of 5 and 15 years;

• Do, if there is a discrepancy between the results from monolinguals vs. bilinguals/multilinguals, typical vs. atypical populations, and adults vs. children, consider whether the task is sensitive enough to capture on-line processing in the particular groups (for a discussion about this issue in relation to L2 learners, see Frenck-Mestre, 2005; 2006);

• Do test all participants with the same equipment or equipment with the same specifications. This is crucial for the way RTs are measured. If these are collected using a button box, then all participants should use the same or the same type of button box. If they are collected with the keyboard or a mouse, then all participants should use the same keyboard or mouse because there can be very large variation between different devices in terms of speed, which can be a confounding factor.
Acknowledgments

The content of this paper is based on my presentations at EMLAR III and EMLAR V. I would like to thank Sharon Unsworth, Elma Blom, and Sergio Baauw for inviting me to present at these two workshops, the audience of the two workshops for providing a stimulating discussion, and two anonymous reviewers for their extremely insightful comments on an earlier draft. This paper was written whilst supported by the ESRC grant RES-061-23-0137.
References


