Transfer of Reading Comprehension Skills to L2 is Linked to Mental Representations of Text and to L2 Working Memory

Catherine Walter
Institute of Education, University of London

Two notions from cognitive psychology were examined in relation to the transfer of reading comprehension skills from L1 to L2: (1) the notion that reading comprehension proceeds by the comprehender’s building of a mental structure representing the text and (2) the notion of working memory. Two groups of French learners of English (at upper-intermediate and lower-intermediate proficiency levels) participated in the study: members of both groups were proficient comprehenders in L1 French, but they differed in their ability to comprehend texts in L2 English, even when the lower-intermediate learners had no problem in processing the individual sentences of those texts. Performance in pro-form resolution in two distance conditions provided strong support for the hypothesis that the lower-intermediate group had failed to transfer to L2 the ability to build well-structured mental representations of texts, while the upper-intermediate group had succeeded in transferring this ability. This structure-building ability was in turn linked to the development of working memory in L2.

INTRODUCTION

The study of the transfer of reading skills from L1 to L2 has a long history. It has sometimes been stated in terms of a threshold effect (articulated by Alderson 1984; Clarke 1988) and studied in several L1–L2 pairs (Barnett 1986; Markham 1985; Bossers 1991; Carrell 1991). Lee and Schallert (1997) addressed the question of whether the phenomenon is better characterized as a continuous one, where the contribution of L1 reading ability to L2 reading ability gradually increases as L2 proficiency increases, or a discontinuous one, typified by the threshold metaphor. While Lee and Schallert’s results were not conclusive, they found more support for a threshold than for a continuous relationship.

All of these studies were prompted by the observation that L2 comprehension skills at lower levels of proficiency do not seem to keep pace with learners’ ability to understand texts sentence-by-sentence. However, none of the studies proposed to define the nature of the comprehension skills that initially fail to transfer and later do transfer.
STRUCTURE BUILDING AND L2 READING COMPREHENSION

The first question that I will examine is this: is the transfer of mental structure-building skill associated with the level of success in L2 reading comprehension?

The idea that comprehension involves the building of coherent mental representations or structures for texts is a familiar one in L1 reading comprehension research (see for example Bransford and Johnson 1972; Carpenter and Just 1977; van Dijk 1977; Johnson-Laird 1983; van Dijk and Kintsch 1983; Schmalhofer and Glavanov 1986; Garnham 1987; Kintsch 1998). In this tradition, I shall adopt Gernsbacher’s (1990) ‘structure building’ model. Gernsbacher gives persuasive empirical evidence that structure building during the comprehension of a narrative involves the same processes whether the narrative is communicated orally, in writing or even as a picture story. In other words, there appears to be an underlying mode-independent structure-building skill, separate from the decoding skills used in listening, reading or interpreting graphic information. It would be parsimonious to suppose that the same structure-building skill underlies comprehension in L2.

In order to judge the evidence for the transfer of the structure-building skill to the L2 reading comprehension context, let us look in more detail at Gernsbacher’s (1990) description of the structure building process:

1 a foundation for a structure is laid at the beginning of a clause or text;
2 subsequent information is mapped onto that foundation so long as the comprehender considers it relevant;
3 this continues in a hierarchical fashion: information relevant to a substructure is mapped onto that substructure;
4 if subsequent information is not considered relevant to an ongoing structure or substructure, then another structure or substructure is initiated.

Note the following points:

• this process is very rapid, and not available to conscious introspection;
• ‘information’ also includes the activation of knowledge in long-term memory;
• there may be more than one ‘good’ structure by which different skilled comprehenders encode the same text;
• the structure in question is not seen as occupying a place in the brain, but is more likely to consist of a spatially dispersed network of neural activation.

In skilled comprehension, readers suppress inappropriate meanings and build large, cohesive, hierarchical structures, in which each new element, as it is integrated, activates preceding related elements. Elements that receive repeated activation are more accessible for recall. In poor comprehension, on the other hand, readers fail to suppress inappropriate meanings and shift
too often to new substructures, building structures that are ‘bulkier, less cohesive and less accessible’ (Gernsbacher 1990: 213).

Thus, if a reader is building an efficient mental structure, a new element is fitted into an appropriate place in the overall framework of that structure, and so activates related elements from earlier in the text. The more frequently an earlier element is activated, the more accessible it remains for recall, so a cohesive structure leads to better recall of earlier elements. In a poor, less-cohesive structure, a new element will activate few previous elements, and earlier elements will be difficult to recall (Gernsbacher 1990: 211–13).

What would this mean for readers who were skilled at structure-building when reading in L1 but had not completely transferred this skill to reading in L2? As skilled comprehenders in an L1 situation, they would recall elements from a text soon after reading them; and building a cohesive structure would also allow them to recall these elements later on. In the L2 situation, if the text was well matched to their level of proficiency, they would process it well sentence-by-sentence, and so would be able to recall elements soon after reading; but building a poor, less cohesive mental structure would cause problems in recalling elements at a distance.

It follows that one way to gauge the skill of a reader in building mental structures is to measure their success in recalling earlier information from the text. Among the different possible indicators of this success, I have chosen co-reference resolution. In co-reference, one expression in the text depends on or presupposes another expression (Brown and Yule 1983: 190–222): examples are pronouns and pro-verbs and their antecedents.

Suppose a group of readers are able to build hierarchical, cohesive mental structures when reading in L1; and suppose that in L2 they are reading a text corresponding to their syntactic and lexical level of proficiency (i.e. a text they can process sentence-by-sentence). Suppose, however, that for some reason they are not able to build sound mental structures based on reading in L2. Gernsbacher’s (1990) framework predicts that these readers will resolve co-refering expressions:

• well in L1, whether the two expressions are near to or remote from one another; and
• well in L2 when the expressions are near to one another; but
• poorly in L2 when the expressions are remote from one another. (If, on the other hand, they are having problems with sentence-by-sentence processing, they will have difficulty resolving co-refering expressions even when they are near to one another.)

In the first part of this study, therefore, I examine whether relative success or failure in structure building in an L2 situation, as measured by co-reference resolution, corresponds to relative success or failure in L2 reading comprehension.
WORKING MEMORY AND READING COMPREHENSION

The second question to be asked in this study is this: is the development of L2 verbal working memory linked to the transfer of reading comprehension skill (and hence structure-building skill) from L1 to L2?

Working memory (henceforth WM) is a system of mechanisms by which humans process the information they need for the performance of complex cognitive tasks and maintain it in an accessible form; that is to say, it is a combination of a processing system and a storage system. What is processed and stored can be information from long-term memory, or new information, or both. Within the general definition of WM, detailed descriptions differ from one model to another, but, as shown in the admirable 1999 volume edited by Miyake and Shah, there is more consensus than disagreement on major issues (Anderson et al. 1996; Baddeley and Hitch 1993; Wilson and Baddeley 1988; Conway and Engle 1994; Cowan 1995; Ericsson and Kintsch 1995; Howes and Young 1996; Kiers and Meyer 1994; O’Reilly et al. 1999; Schneider 1999; Teasdale and Barnard 1993).

There is ample support in the literature for the involvement of WM in L1 reading comprehension (Daneman and Carpenter 1980; Baddeley et al. 1985; Daneman and Tardif 1987; Engle et al. 1992; Cantor and Engle 1993; Ericsson and Delaney 1999). In the studies of Yuill and Oakhill (1991) in the UK and Cornoldi et al. in Italy (1996) comparing groups of children who were either good L1 comprehenders, or good decoders but poor comprehenders, the researchers found consistently lower WM capacity for the poor comprehenders. Swanson and Berninger (1995) studied 91 skilled and less-skilled child readers in English-speaking Canada and found that low WM (but not low phonological short-term memory) correlated with poor reading comprehension, supporting the hypothesis that WM deficits contribute to reading comprehension problems. These are only a few examples of a large body of literature (see also Whitney et al. 1991; Oakhill and Yuill 1996; Perfetti et al. 1996; Light and Capps 1986).

In L2, Geva and Ryan (1993) studied 10- to 12-year-old English–Hebrew bilinguals and found that their performance in reading comprehension in L2 was accurately predicted by performance on static and working memory storage tests, in combination with ‘underlying intelligence and L2 oral proficiency’. Their subjects had learnt to read in L1 at the same time as they were learning both to speak and to read in L2. Geva and Ryan did not find any effect of years of L2 instruction on L2 reading comprehension. However, it is difficult to compare this study with others focusing on the transfer of reading comprehension skills from L1 to L2, since the latter consider individuals who have learnt to read proficiently in L1 before beginning L2 instruction.

Also in L2, Harrington (1992) and Harrington and Sawyer (1992) studied WM in Japanese–English bilingual university students, and found a strong correlation between their L2 WM measure and reading comprehension as
measured by the TOEFL. Note, however, that this work focused on learners who were classified as bilingual.

The focus of the present study is the transfer of reading comprehension skills from L1 to L2, at a much lower level of L2 proficiency than the bilinguals in the Geva and Ryan or Harrington and Sawyer studies.

**METHODOLOGY**

**Participants**

In order to control the number of variables in the study, two groups of participants were selected from the same L1 cultural and educational background. Choosing to keep these linguistic and sociological variables constant while varying the level of L2 proficiency meant that the mean age and the educational attainment of the two groups were different, and this can be seen as a weakness in the design. The older, more academically advanced group is virtually guaranteed to comprehend somewhat more skillfully than the less advanced group, even in L1, and this needed to be taken into account in the analysis of results. However, the alternatives (selecting groups from different L1, educational or cultural backgrounds in order to control for age and/or educational attainment) would almost certainly have led to even more serious threats to the design.

The participants were French native speakers from a middle school (Collège) and an upper school (Lycée) in a provincial town in the Savoie department of France. They came from a range of socioeconomic categories and lived either in the town or in one of the nearby mountain resorts. Four possible candidates were excluded after initial interviews, since they were bilingual from birth and spoke both languages at home. Thus all participants in the study were native speakers of French from monolingual households, who did not speak any other language on a regular basis outside of a classroom setting. Two other participants (one from each of the two groups) were excluded when they left school for psychosocial reasons near the beginning of the study. This left two groups of 19 and 22 participants respectively.

The lower-intermediate group of participants consisted of 19 adolescents (13 girls and 6 boys, mean age 14;6 years, with a range of 13;5 to 15;11 years) from a single English class in the final year of Collège (middle school). A wide range within a year group is not unusual in the French school system, where little stigma is attached to repeating a year (Royal 1999). The class was of mixed ability but it did not include any pupils with special educational needs. They were in their fourth year of English classes (3 × 50 minutes a week), and in their second year of another foreign language (usually German or Italian).

The upper-intermediate group consisted of 22 young people (18 young women and 4 young men, mean age 17;10 years with a range of 16;11 to 19;6 years) from a single English class in the final year of Lycée (upper school).
Only pupils with sufficiently high marks on the examinations at the end of middle school are admitted into upper school, so overall academic ability in this group will have been more uniform than in the younger group. The upper-intermediate participants were in their seventh year of English classes, and had studied or were studying at least one other foreign language.

Participants also took part in three other experiments besides those described and referred to here; each participant received a present at Christmas (a book for upper-intermediate participants and a bag of sweets for lower-intermediate participants) and a book voucher at the end of the series of experiments.

Materials and procedures

Three measures were administered, each in both languages: a baseline comprehension assessment, a pro-form resolution test and a working memory measure.

Baseline comprehension measure

The baseline comprehension measure was undertaken to establish that both groups were skilled comprehenders in L1 and to confirm that a difference in L2 comprehension skill could not be attributed to decoding or sentence-level processing differences. For this measure, a summary completion task, that is a gapped summary to complete, was chosen. Taylor (1996: 83–8) argues persuasively that summary completion offers a reading comprehension assessment format which takes account of readers’ natural processing of a text. This distinguishes it from multiple choice questions or discrete-item comprehension questions, whose basis in an analysis of reading as a set of subskills has been called into serious question (by, for example, Alderson 2000: 93–7). Moreover, summary completion, unlike question-answering, does not require the reader to understand additional material besides that which is already included in the text: a particularly important point in L2 comprehension testing.

The English language stimuli were taken from graded readers designed for learners of English as a foreign language, chosen from the list of the 600 books judged to be the best in print by the Edinburgh Project for Extended Reading (EPER 1998). Short (124- to 150-word) self-contained texts were chosen from sixteen graded readers which corresponded to the proficiency level in English of the lower-intermediate group as determined by (1) the EPER level; (2) their teacher’s assessment of the texts, (3) comparison with their course book, and (4) observation of the class. Most texts were from the very beginning of each book. Texts were chosen for similarities in style and topics as well as in level and length. Some stories focused on action and plot and some on exploring feelings and relationships, so as not to favour boys’ or girls’ preferred narrative styles (Millard 1997). Care was taken that the stories did not depend on
cultural schemata that the French participants would not share (Steffensen et al. 1979). The sixteen French texts were translations of the English texts, checked for naturalness by two French native speaker linguists.

As a final guarantee of the accessibility of the English texts for the lower-intermediate group, the content, vocabulary, and grammatical structures in the English versions of the stories were checked by their English class teacher, who felt comfortable that the children would not have any significant difficulty with these. Additionally, she included all less frequent words for productive use in class work shortly before the trials.

For each text, a straightforward summary of about fifty words was written, containing five gaps. Seven native English-speaking volunteers (three linguists and four adolescents of the same age as the lower-intermediate participants) attempted to complete the English summaries without reading the original texts, in order to make certain that no answers were unequivocally internally predictable. The adolescents were paid £5 an hour.

The texts and their summaries were entered into the Psychscope programme (Cohen et al. 1993) on a Macintosh PowerBook.¹

The languages of the texts were counterbalanced, so that half of the participants read each text in one language and half of the participants read it in the other language. The order of the texts was rotated. To balance the risks of other-language activation,² familiarity effect, and fatigue effect, the language was changed every two texts: two texts in one language were followed by two texts in the other language and so on. Each text was preceded first by a screen with a row of asterisks, then by a screen stating the language of the following text (‘English’/‘français’), and finally by a screen with the title of the text.

Each summary appeared on a single screen after its text, from which it was separated by a screen with a row of asterisks and then a screen with the title of the text and instructions in the appropriate language. Paper answer sheets gave an entire line for each blank in the summary. A sample text and its summary are given in Appendix 1.

Piloting was conducted in English with nine native English-speaking girls (mean age 13:7 years) and in both languages with eleven French learners of English (mean age 17:2 years). The aims of this piloting were (1) to confirm that the basic stories were clear and interesting to this age group; (2) to check the summaries for appropriate difficulty and for clarity (answers were discussed with each volunteer after the trial); (3) to use the answers in order to test and fine-tune the marking system; and (4) to confirm that the texts presented no cultural difficulties for French participants.

In the study, trials were conducted with participants individually. After receiving a spoken explanation in French and before beginning the trials, each participant practised self-paced reading on the computer with a story in French; and filled out a sample answer sheet. The participants were told that neither spelling nor grammar was important in filling out the summaries; that they could put as many or as few words as they wished for each blank; and
that they could write answers in French for blanks in English summaries if they so wished. ‘Meaning is all that counts’ was the final instruction for summary completion.

The participant pressed a key to move from one screen to the next, but could not go back to a previous screen. If a screen was skipped because of faulty handling of the computer, however, the experimenter told the participant what had been on the skipped screen or showed it to them on paper (this happened three times). As a further check that decoding or sentence-level processing problems did not interfere with comprehension, participants were invited to ask the meaning of any words they did not understand in English, either in the text or in the summary. No participant asked about more than a total of six words for all of the eight English stories, and the questions were overwhelmingly to confirm correct knowledge, for example, ‘Window, c’est fenêtre?’.

In scoring, one point was awarded for each of the five gaps in each summary. Half-points were allowed for answers which, while not complete, gave an indication that the participant had a partial understanding of the information required. Criteria for scoring had been established on the basis of the piloting operation. They were then adjusted using the first ten answer leaflets from each group as feedback, that is a few additional acceptable answers occurred in the first ten leaflets, and these answers were added to the list of acceptable answers. These twenty leaflets were then re-marked, and all the subsequent leaflets were marked using the same criteria.

Pro-form resolution measure

As in Yuill and Oakhill’s (1991: 85–95) study of young L1 comprehenders’ resolution of co-referring expressions, the co-referring expressions were embedded in naturalistic stories. This procedure was chosen rather than using several passages which would require participants to set up a new context for each pro-form.

Two stories (‘Burglars’, 875 French/865 English words; and ‘McCarthy’, 951 French/938 English words) were written. The stories contained both action and exploration of feelings, in order not to privilege girls’ or boys’ reading styles (Millard 1997). To ensure that the English texts did not present decoding or sentence-level processing difficulties for the lower-intermediate group, the same procedures were followed as for the comprehension assessment. To check for naturalness of discourse, five educated native speakers with no training in linguistics were asked to read the story and say ‘if there was anything strange about the way they were written’; three English-speaking and two French-speaking linguists were asked to read the story and say if the discourse seemed to proceed unnaturally at any point. No problems were signalled by any of the readers.

Instances of pronoun and pro-verb co-reference were chosen (16 for the ‘Burglars’ story and 15 for the ‘McCarthy’ story) and when the stories were
printed on paper these words were printed in red to contrast with the ordinary black typeface of the rest of the story. The beginning of the ‘Burglars’ story is given in Appendix 3.

The experimenter and three other linguists made independent judgements about the pro-forms in the two stories, putting them into one of two distance categories:

- immediate, where the first mention is in the same or previous clause;
- remote, where the first mention is more than two clauses distant, with any clear mediating mention also two or more clauses distant.

In each of the two stories, ten pro-forms for which all four linguists agreed on classification were used for scoring (four pronouns and one pro-verb in each distance condition). Thus, in each language, each participant was scored on the resolution of five pro-forms for each distance condition (immediate, remote). There was always another plausible first mention in the story for the scored item.

Trials were conducted with participants individually. After an explanation in French and a practice session in French and English on several two-sentence passages, the participants each silently read one story on paper in each language. They were told to stop when they got to a word in red, read the word aloud, and tell the experimenter ‘what the word meant’, in either language, in as few or as many words as they wished. Half the participants in each group read ‘Burglars’ in French and ‘McCarthy’ in English, and the other half read with the languages reversed. Within these groups, half the participants read the French story first and half read the English story first. When they encountered an expression printed in red, they read the word aloud and then tried to identify the first mention. If the response was incorrect, the experimenter first indicated merely that something was not right; if this did not elicit the correct answer the experimenter gave some guidance, recalling the context of the story. This procedure was chosen for several reasons:

- to allow for the possibility of marking the results either strictly or liberally, to avoid ceiling or floor effects (in the event, however, only the spontaneous responses were used in the scoring)
- to prevent an early error from influencing later items. The decision had been made to embed the trial items in one story for each language rather than writing a new mini-story and obliging participants to set up a new context for each item. However, a risk of this approach is that earlier errors can influence later decisions; this risk was minimized by the procedure described. Note that the instrument was measuring pro-form resolution skill, and not reading comprehension, which had been assessed in the baseline study.
- to put the participants at their ease, and to avoid their leaving the trials with a sense of failure.
As a further check that decoding or sentence-level processing problems did not interfere with comprehension, participants were also encouraged to ask for the meaning of any words or expressions that they did not understand in English. There were very few problems with lexis, and participants’ questions were overwhelmingly confirming in nature.

A correct answer given without any prompting was scored 1 and all other responses were scored 0. (More liberal scoring did not give different results.)

Verbal working memory measure

The verbal WM measure used in this study is fully described in Walter (2000) and Walter and Williams (in preparation). This measure is based on Waters and Caplan’s (1996) modification of the classical Daneman and Carpenter (1980) reading span measure. The great advantage of a measure of this type over the Daneman and Carpenter measure is this: WM comprises both processing and storage; but while the Daneman and Carpenter measure requires subjects both to process and to store, it only measures storage. Like Waters and Caplan’s (1996) instrument, the WM measure in the present study measures both processing and storage.

In each language, the stimuli consisted of sets of simple declarative sentences, six to eight words long, in which the argument requirements of the verb are either respected (‘logical’ sentences) or not (‘illogical’ sentences). That is to say, in half the sentences (‘illogical’ sentences) an inanimate subject is used with a verb which requires an animate one (e.g. Several of the flutes play my cousins). The sentence-final words are all different concrete nouns. Sentences were randomly ordered for ‘logical’/‘illogical’ conditions. They were then divided into four series of five increasingly longer sets of sentences: the first series contained five two-sentence sets and the last series contained five five-sentence sets. Adjustments were then made (respecting the ‘logical’/‘illogical’ order) to ensure that, in any set, the sentence-final words were not phonologically similar and did not have obvious semantic links. There were four versions of the materials, varying along two parameters:

1. In each language, the original randomized list was reversed and re-divided into sets, and small adjustments were made to avoid phonological similarities and semantic links.
2. Two of the versions began with a series of five two-sentence French sets and two began with a two-sentence English series. In all versions, the language was changed after the initial two-sentence series and then after every two series.

The sentences were entered into the Psyscope programme (Cohen, Mac-Whinney, Flatt, and Provost 1993) on a Macintosh PowerBook, to appear in the centre of the screen, one sentence at a time. Each set of sentences was followed by a screen showing a line of asterisks. Each group of five sets was preceded by a screen announcing the language and length of the next set.
Participants judged whether each sentence was logical or illogical (processing measure); the time taken for this judgement was recorded (processing measure); and, at the end of a set, participants attempted to remember the last word of each sentence in order (recall measure). Each participant was allowed to choose which of 1 or 0 on the computer keyboard they preferred for ‘logical’ and ‘illogical’ choices, and practised with sample material until they were at ease with the procedure. Pressing the button to signal ‘logical’ or ‘illogical’ made the next sentence or a row of asterisks appear on the screen. When the asterisks appeared, the participant attempted to say the sentence-final words of the set in order and the result was recorded on paper by the experimenter. Success in an item was defined as correct recall of all sentence-final words in a set, in order, with this exception: if a participant said the sentence-final words correctly but in the wrong order, the experimenter asked (depending on the language in use) ‘Dans l’ordre?/In order?’, and a correct response to this prompt was counted as a success. The experimenter did not otherwise indicate whether individual responses were right or wrong. The trial continued until the participant could no longer reliably recall final words.

Working memory scores in each language were calculated for each participant, based on three figures: one measure of storage (reliably remembered sentence-final words, or ‘recall span’), and two measures of processing (reaction time in milliseconds for correct answers at recall span; and percentage of correct logicality judgements).

Recall span, as in Daneman and Carpenter (1980) and Waters and Caplan (1996), was expressed as the highest level (sentences per set) at which the participant correctly recalled the sentence-final words in three out of five sets, with an extra half-point given for correct recall in two of the five sets at the next higher level. For example, a participant who recalled all sentence-final words correctly on three out of five four-sentence sets, and then recalled all sentence-final words correctly on two out of five five-sentence sets, would obtain a recall span score of 4.5.

Reaction time was measured as the time in milliseconds between the appearance of a sentence on the screen and the pressing of the button to signal the logicality judgement. Only reaction times at reading span for correct logicality judgements were taken into account.

To obtain an overall WM score, z-scores for the three components (recall span, percentage of correct logicality judgements, reaction time) were calculated by taking both French and English components into account. In other words, to take recall span as an example, each subject contributed two values (recall span in French and recall span in English) to the sample population on which the recall span z-scores were based. The z-scores for reaction time at span and percentage of correct logicality judgements were calculated in the same way. The reaction time z-scores were multiplied by –1 (higher reaction times indicating less good processing, whereas higher scores on the other two measures indicate better processing and recall). Overall WM scores were calculated as the means of the three z-scores.
RESULTS

Baseline reading comprehension assessment
Mean scores out of 40 and standard deviations are given in Table 1. Figure 1 shows the summary completion scores by language and group.

From Table 1 and Figure 1, showing the results of the baseline reading comprehension assessment, it can be seen that in L1 French, while both groups showed good evidence of comprehension, at 80 per cent and 70 per cent of correct summary completion replies, there was a small difference between them. This difference was confirmed by a one-way ANOVA in which there was an effect of group, $F(1,39) = 7.041, p < .05$. This difference was not surprising, given the disparity in maturity and academic attainment between the two groups. However, at 70 per cent of correct judgements, the lower-intermediate participants were comprehending well in L1. The next question was whether the between-group difference in comprehension scores in L2 paralleled the difference in L1; or whether the lower-intermediate group’s L2

Table 1: Means and standard deviations of summary completion scores for all participants (out of 40), by language and group

<table>
<thead>
<tr>
<th>Language</th>
<th>Lower-intermediate (N = 19)</th>
<th>Upper-intermediate (N = 22)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SD</td>
</tr>
<tr>
<td>French</td>
<td>28.07</td>
<td>4.18</td>
</tr>
<tr>
<td>English</td>
<td>13.84</td>
<td>7.21</td>
</tr>
</tbody>
</table>

Figure 1: Mean summary completion scores (out of 40), by language and group
scores were much further from their L1 scores than their upper-intermediate counterparts’ scores.

In order to investigate whether there was a difference in the degree to which the two groups were transferring their comprehension skills from L1-based comprehension to L2-based comprehension, a 2×2 mixed-design ANOVA was carried out on the summary completion scores, with one between-subjects factor: group (upper-intermediate, lower-intermediate), and one within-subjects factor: language (French, English). Both main effects were significant. There was a main effect of group, with upper-intermediate participants completing summaries more successfully than lower-intermediate participants: F(1,39) = 47.20, p < .0001. There was a main effect of language, with summaries of French stories being completed more successfully than summaries of English stories: F(1,39) = 59.70, p < .0001.

There was a significant two-way interaction between language and group: F(1,39) = 17.95, p = .0001. This demonstrates that the lower-intermediate participants’ comprehension scores in the two languages were much further apart than those of their upper-intermediate counterparts. (Stricter and more liberal scoring regimes gave the same results.)

Pearson product–moment correlations were carried out between the mean French and English summary completion scores, to investigate whether French L1 comprehension skill predicted English L2 comprehension skill. There was an overall correlation between the two sets of summary completion scores, r = .31, p < .05, but no significant correlations when the scores were broken down by group. Pearson product–moment correlations were also carried out between the English mean summary completion scores and English final school marks from the previous academic year as based on the schools’ proficiency examinations. Here, in contrast, there were correlations not only for both groups taken together (r = .39, p < .05), but also for each group taken on its own (lower-intermediate: r = .73, p < .001; upper-intermediate: r = .61, p < .01). Note that these school marks were based on different examinations for each group, but that the examinations were of the same nature and based on the same scales. Within each group, proficiency in English, as measured by school marks, predicted success in English summary completion better than did L1 French reading comprehension skill measured as success in summary completion; this pattern is current in the literature about reading at lower levels of L2 proficiency (in, for example, Bernhardt and Kamil 1995; Carrell 1991).

Recall that the comprehension texts were specifically written and tested so that the lower-intermediate group could process the language at a sentence level. Yet the difference between lower-intermediate L1 and L2 text comprehension scores (36 percentage points) was significantly larger than the upper-intermediate L1/L2 difference (10 percentage points). It is unlikely that the difference between the L1 scores of the two groups would cause such a large disparity between their L2 scores without the intervention of some other factor. It can thus be said with confidence that these two groups of
readers were (a) processing well at sentence level and comprehending well at text level in L1; (b) processing well at sentence level in L2; and (c) showing patterns of transfer (upper-intermediates) or difficulty in transfer (lower-intermediates) of reading comprehension skills from L1 to L2 texts, similar to the patterns found in previous studies.

Pro-form resolution measure

Mean scores out of 5 and standard deviations are given in Table 2. Figure 2 shows the pro-form resolution scores by language, distance, and group.

<table>
<thead>
<tr>
<th>Language</th>
<th>Distance</th>
<th>Lower-intermediate (N = 19)</th>
<th>Upper-intermediate (N = 22)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Mean</td>
<td>SD</td>
</tr>
<tr>
<td>French</td>
<td>Immediate</td>
<td>4.10</td>
<td>0.87</td>
</tr>
<tr>
<td></td>
<td>Remote</td>
<td>3.73</td>
<td>1.03</td>
</tr>
<tr>
<td>English</td>
<td>Immediate</td>
<td>3.74</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>Remote</td>
<td>2.26</td>
<td>1.33</td>
</tr>
</tbody>
</table>

*Figure 2: Pro-form resolution scores (out of 5), by language, distance, and group*
A mixed-design $2 \times 2 \times 2$ ANOVA was carried out, with one between-subjects factor: group (upper-intermediate, lower-intermediate), and two within-subjects factors: language (French, English) and distance (Immediate, Remote). All main effects were significant: a main effect for group, $F(1,39) = 10.58, p < .005$, indicating that the upper-intermediate participants resolved pro-forms more successfully than the lower-intermediate participants; a main effect for language, $F(1,39) = 8.53, p < .01$ indicating that L1 pro-forms were resolved more successfully than L2 pro-forms; and a main effect for distance, $F(1,39) = 27.47, p < .0001$, indicating that Immediate pro-forms were resolved more successfully than Remote pro-forms. There was a significant two-way interaction between language and group, $F(1,39) = 6.39, p < .05$, indicating that the upper-intermediate participants’ scores in L2 were reliably closer to their scores in L1 than those of their lower-intermediate counterparts. There was a significant two-way interaction between distance and group, $F(1,39) = 4.33, p < .05$, indicating that the upper-intermediate participants’ scores for Remote pro-forms were reliably closer to their scores for Immediate pro-forms than those of their lower-intermediate counterparts. Most importantly, there was a significant three-way interaction between language, distance and group, $F(1,39) = 6.91, p < .05$, indicating that the difference in L1 and L2 pro-form resolution between the two groups in the Remote condition was reliably different from that in the Immediate condition. (Note that an ANOVA with ‘Assignment of story version’ nested within group was carried out, but its results were identical to those of the $2 \times 2$ ANOVAs reported here.)

Pearson product–moment correlations were performed between L2 baseline comprehension and L2 resolution of Remote pro-forms for each group. The results are given in Table 3.

The upper-intermediate participants were resolving Immediate and Remote pro-forms similarly well in L1 and in L2. The lower-intermediate participants resolved Immediate and Remote pro-forms well in L1, and resolved Immediate pro-forms well in L2. This is consistent with their building effective structures corresponding to the L1 text, and with their processing the L2 text well at sentence level. However, the lower-intermediates were significantly worse at resolving Remote pro-forms in L2 than in L1. If they had experienced difficulty in sentence-level processing in L2, this would have been reflected in their Immediate L2 pro-form resolution scores, and this was not the case. The patterns of the lower-intermediate scores correspond to what was predicted if they were building effective structures corresponding to L1 texts, but failing to build effective structures corresponding to L2 texts. In other words, these results support the hypothesis that the transfer of mental structure-building skill is associated with the level of success in L2 reading comprehension.

In the baseline comprehension measure, we saw two groups of learners at two different levels of L2 proficiency. The upper-intermediate group showed via the summary completion measure that they comprehend reading texts similarly well in L1 and in L2. The lower-intermediate group results in
summary completion showed that they comprehend well in L1; but that they have problems comprehending L2 reading texts, even when these are geared to their level of L2 proficiency. Now we see that, in the Remote pro-form resolution measure, designed to gauge structure building skill, the lower-intermediate group performs well in L1 and poorly in L2. This is consistent with success in transfer of structure building skill being associated with success in L2 comprehension.

In this regard, the pattern of correlations in Table 3 is interesting. There are no correlations between summary completion and Remote pro-form resolution in L1, and this may indicate that both groups were sufficiently skilled in structure building in L1 so that individual differences were not reflected. In L2, the lack of correlation for the lower-intermediate group is unsurprising given their relative lack of success in both measures. What is interesting is the significant correlation in L2 for the upper-intermediate group, whose members were successful in L2 in both comprehension and summary building measures. The better an upper-intermediate participant was at structure building, the better s/he was at comprehension. This provides further support for the hypothesis that the transfer of structure building skill is associated with success in L2 comprehension.

Table 3: Pearson product-moment correlations between summary completion scores and remote pro-form resolution scores, as a function of language and group

<table>
<thead>
<tr>
<th></th>
<th>Lower-intermediate (N = 19)</th>
<th>Upper-intermediate (N = 22)</th>
<th>Total (N = 41)</th>
</tr>
</thead>
<tbody>
<tr>
<td>French L1</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
</tr>
<tr>
<td>English L2</td>
<td>ns</td>
<td>.63**</td>
<td>.65***</td>
</tr>
</tbody>
</table>

** p < .01; *** p < .0001

Working memory measure

Means and standard deviations for percentage of correct logicality judgements, number of final words reliably recalled and reaction time for correct answers at best recall span by language and by group are given in Table 4.

The means and standard deviations of overall WM scores, by language and group, are given in Table 5.

In order to examine the overall relationship of WM with the baseline comprehension (summary completion) scores, both French and English WM scores were plotted on the x-axis of the same graph, against the means of the corresponding summary completion scores. The resulting correlations are given in Table 6. Correlations were compared using Fisher’s (1921) method for testing the difference between two independent r's, and this showed
Table 4: Means and standard deviations of scores for recall span (maximum 5.0), logicality judgement (percentage of correct judgements) and reaction time for correct answers (in milliseconds, at span) for all participants, by group

<table>
<thead>
<tr>
<th>Measure</th>
<th>Language</th>
<th>Lower-intermediate (N = 19)</th>
<th>Upper-intermediate (N = 22)</th>
<th>Total (N = 41)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Mean</td>
<td>SD</td>
<td>Mean</td>
</tr>
<tr>
<td>Recall span</td>
<td>French</td>
<td>3.73</td>
<td>0.79</td>
<td>4.50</td>
</tr>
<tr>
<td></td>
<td>English</td>
<td>2.05</td>
<td>0.95</td>
<td>2.84</td>
</tr>
<tr>
<td>Logicality</td>
<td>French</td>
<td>94</td>
<td>4</td>
<td>96</td>
</tr>
<tr>
<td>judgement</td>
<td>English</td>
<td>83</td>
<td>12</td>
<td>91</td>
</tr>
<tr>
<td>Reaction time</td>
<td>French</td>
<td>6144</td>
<td>1615</td>
<td>5278</td>
</tr>
<tr>
<td></td>
<td>English</td>
<td>11785</td>
<td>3704</td>
<td>6522</td>
</tr>
</tbody>
</table>

Table 5: Means and standard deviations of WM scores (expressed as means of the z-scores of the three contributing measures, with the reaction time z-score polarities reversed) for all participants, by language and group

<table>
<thead>
<tr>
<th>Language</th>
<th>Lower-intermediate (N = 19)</th>
<th>Upper-intermediate (N = 22)</th>
<th>Total (N = 41)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SD</td>
<td>Mean</td>
</tr>
<tr>
<td>French</td>
<td>0.35</td>
<td>0.25</td>
<td>0.73</td>
</tr>
<tr>
<td>English</td>
<td>−1.01</td>
<td>0.59</td>
<td>−0.02</td>
</tr>
</tbody>
</table>

Table 6: Correlations between all WM scores and summary completion scores

<table>
<thead>
<tr>
<th>All scores</th>
<th>French scores</th>
<th>English scores</th>
<th>Upper-intermediate</th>
<th>Lower-intermediate</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>.76**</td>
<td>.33*</td>
<td>.73***</td>
<td>.46**</td>
</tr>
</tbody>
</table>

* p < .05; ** p < .01; *** p < .0001

significant differences between the French and English correlations (p < .05) and the lower-intermediate and upper-intermediate correlations (p < .01).

In other words, Table 6 shows that having a higher verbal WM (ability to process and store complex information simultaneously) corresponds to being
better at reading comprehension, for both groups. It also shows that this is significantly more the case for the lower-intermediate group. Small advantages in WM give significant advantages in comprehension, especially at the lower level of proficiency. This supports the hypothesis that there is a link between the development of verbal WM in L2 and success in L2 reading comprehension.

When correlations were carried out between WM and summary completion scores for each language by group, neither group yielded a significant correlation in French, and nor did the upper-intermediate group yield a correlation in English. The only significant correlation in English was between the lower-intermediate group WM scores and their baseline comprehension scores, \( r = 0.49, p < .05 \).

This single significant correlation is similar to findings in other studies of WM. Observation of WM differentiation ‘when the going gets tough’ is common in the literature, and is well accounted for in various models (e.g. Baddeley and Logie 1999; Ericsson and Kintsch 1995; Ericsson and Delaney 1999; Engle et al. 1999). Neurophysiological support for this observation is provided by D’Esposito et al. (1995), who performed fMRI studies on patients who were simultaneously carrying out a spatial task and a verbal task; D’Esposito’s group found that the dorsolateral prefrontal cortex—the part of the brain that has been found to be associated with WM executive control processes—was activated (as were the domain-specific areas associated with the two tasks). It is notable that the prefrontal cortex was not measurably activated when the two tasks were performed independently, but only when there was a higher overall demand on WM. In the present study, the correlation between WM scores and comprehension scores corresponds well to the lower-intermediate group’s experiencing high demand on their L2 WM in the summary completion tasks.

In order to examine the relationship of verbal working memory with structure building, both French and English WM scores were plotted on the x-axis of the same graph, against the means of the corresponding Remote pro-form resolution scores. The resulting correlations are given in Table 7. Detailed correlations between English WM and English Remote pro-form resolution scores by group were performed, but these did not, however, yield significant correlations.

The Table 7 figures do tend to support the hypothesis of higher demands being made on WM in the English pro-form resolution condition. They also correspond to higher WM demands on the lower-intermediate group generally. However, the predicted correlation between Remote pro-form resolution and WM for the lower-intermediate group did not emerge, and this merits further study.
Table 7: Correlations between all WM scores and Remote pro-form resolution scores

<table>
<thead>
<tr>
<th>All scores</th>
<th>French scores</th>
<th>English scores</th>
<th>Upper-intermediate</th>
<th>Lower-intermediate</th>
</tr>
</thead>
<tbody>
<tr>
<td>.45**</td>
<td>ns</td>
<td>.52*</td>
<td>ns</td>
<td>.53**</td>
</tr>
</tbody>
</table>

* p < .001; ** p < .0001

DISCUSSION

The first hypothesis that this study addressed was:

The transfer of mental structure-building skill is associated with level of success in L2 reading comprehension.

This hypothesis was supported: the baseline reading comprehension measure showed that both groups of participants were skilled comprehenders in L1, but differed in their ability to transfer their comprehension skill to L2. Next, the pro-form resolution measure showed that both groups were building sound mental structures in L1, but that there was a significant difference between the groups’ ability to do so in L2, even when they understood texts on a sentence-by-sentence basis. Patterns of correlation supported the notion that the more successful structure builders were operating similarly in L1 and L2, and that the better they were at L2 structure building, the better they were at L2 reading comprehension.

The second hypothesis was:

The transfer of reading comprehension skill (and hence structure-building skill) from L1 to L2 is linked to the development of verbal working memory in L2.

This hypothesis also received some support: verbal WM correlated with the baseline reading comprehension measure precisely in that ‘tough’ circumstance (lower-intermediate, L2 Remote pro-form resolution) predicted by models of WM. In addition, L2 verbal working memory tended to correlate with L2 structure building as measured by Remote pro-form resolution. However, there was not a significant correlation between Remote pro-form resolution and WM for the lower-intermediate group specifically, and this deserves further study.

So a crucial element in transferring reading comprehension skill from L1 to L2 appears to be the transfer of the structure-building ability; and this transfer appears to be linked to the development of WM in L2. These findings provide a possible explanation for the perceived threshold nature of the transfer. If successful structure building is accomplished in L1 but not in L2, it is not the ability to build mental structures that is absent; what is lacking is the
attainment of some level of L2 ability which acts as a precondition for the structure-building skill to operate. Recall the nature of the structure-building skill as described by Gernsbacher (1990). If some element or set of elements in L2 WM were insufficiently developed for the comprehender to link incoming material to a previously established structure, the comprehender would fall back on initiating a new structure. The result of this process occurring repeatedly would be the sort of bulky, inaccessible mental representation that less-skilled L1 comprehenders produce (Gernsbacher, 1990). A failure to link appropriately even once or twice at crucial early points in the building of a mental representation might be enough to prevent the building of a cohesive structure, and hence to prevent recall of earlier information. The L2 comprehender would continue to have difficulties in recall for a time, even as the underlying capacity continued to develop progressively. Success in building reliable mental representations based on L2 would only improve when the underlying capacity had developed to a point where quite a large number of the unconscious decisions involved in building a representation were successful. This would account for the perceived threshold nature of the L1-to-L2 transfer.

The question now arises: what is the nature of the link between L2 WM and L2 reading comprehension? One way of investigating this question is to examine empirically the possible sources of demand on WM capacity, for example less well elaborated L2 phonological representations, or more effortful syntactic processing. Another is to undertake a longitudinal study in which WM, L2 reading comprehension skill, and various elements of L2 proficiency are tracked. This would eliminate the problems associated with any experimental design that involves matched groups, more pronounced in the present case where it is very unlikely that groups can be matched both for first language background and for age.

Further studies involving learners from different L1 backgrounds, including languages outside the Indo-European family and languages whose orthography is not phonological, would throw additional light on this subject.

(Final version accepted April 2004)

ACKNOWLEDGEMENTS

I am most grateful to the participants from the Collège Jean Rostand and the Lycée Régional Polyvalent et Professionel, both in Moûtiers, France; to Michael Swan and John Williams for their continuing advice and support; to Julie Norton, Susan Mandala, Dominique Flandin-Granget, and Catherine Cobés for serving as linguist informants; to the members of the Working Memory Workshop, and especially Alan Baddeley and Robert Logie, for their advice on working memory matters; and to five anonymous reviewers for their very helpful comments on earlier versions of this article.
APPENDIX 1

Project Omega

Charles Hatfield Baker III was a very rich man. There wasn’t a richer man in all of New
York City. He was the head of a large number of important companies. He had land
and buildings and houses in the country. He had boats and horses and an aeroplane.
But he was a simple man. He liked the simple things of life: a good cup of coffee, a kiss
from a pretty girl, a walk through the park on a fine autumn day.

He had all these things on this Wednesday morning. He drank two cups of very good
Italian coffee for breakfast. His daughter, Julia kissed him goodbye. ‘Have a nice day,
daddy,’ she said. And he left his apartment to walk to his office. He always walked
through Central Park to his office.

(Reprinted by permission of Addison Wesley Longman Limited from Project Omega by
Elaine O’Reilly, 1991.)

Summary

Charles Hatfield Baker III was ____A____ New York City. He controlled a lot
of ____B____. He had land, buildings, houses, and expensive possessions. He
liked ____C____ things. On this Wednesday he had two ____D____, his
daughter kissed him, and then he left his apartment to ____E____.

APPENDIX 2

(The beginning of one of the stories from the pro-form resolution test. The
words which are underlined here were printed in red.)

Burglars—English

It was a very difficult day at the office. Monica’s secretary was ill, and a temp
from an agency was taking her place. She wasn’t very efficient, and Monica
had to explain the same things to her again and again. Then there was a
problem in another part of the building that Monica had to go and sort out—
she was away from her desk for more than an hour, and while she was out
she missed an important phone call. [. . .]
NOTES

1 The texts were entered into the Psycscope program so as to appear in one of four segmentation conditions: sentence-by-sentence; word-by-word; by meaningful phrases; or cut in the middle of syntactic units. This was done to test hypotheses about possible differences in reading performance based on segmentation condition. In the event, there was no significant difference between any of the conditions.

2 To some extent, both languages are always activated in bilinguals. But excessive other-language activation can interfere significantly with processing in tasks where the participants have no indication of which language is going to come next. However, studies have shown that bilinguals can exercise a great deal of conscious control over excessive other-language activation, especially when it is clearly indicated which language is called upon for each part of a task, and also when languages are not switched too often. (See Scarborough et al. 1984; Durgunoglu and Roediger 1987; Basden et al., 1994; Costermans and Galland 1980; Doctor and Klein 1992; and Durgunoglu and Hancin’s 1992 review of the subject.)

REFERENCES


Fisher, R. A. 1921. ‘On the probable error of a coefficient of correlation deduced from a small sample.’ Metron 1: 3–32.


Geva, E. and E. B. Ryan. 1993. ‘Linguistic and cognitive correlates of academic skills in first
and second languages.’ Language Learning 43: 5-42


Swanson, H. L. and V. Berninger. 1995. ‘The role of working memory in skilled and less


