

Psicológica (1999) 20, 65-90.

Psicología de la lectura y procesamiento léxico visual: Una revisión de técnicas experimentales y de procedimientos de análisis

Manuel Perea* y Eva Rosa

Universitat de València

Este trabajo ofrece una revisión crítica de las técnicas más frecuentemente empleadas en la investigación psicolingüística, así como de los procedimientos de selección de materiales y análisis de datos en esta área de conocimiento. En la sección correspondiente a las técnicas, el énfasis se pone en los procedimientos para estudiar el reconocimiento visual de palabras y en los procesos subyacentes a los mismos (ya sean comunes a la lectura o específicos de la tarea experimental). En la segunda parte del trabajo, el interés pasa —tras un análisis de la lógica del contraste de hipótesis— a los procedimientos de análisis de la variable dependiente de mayor interés en psicología del lenguaje, el tiempo de reacción. Finalmente, se estudia el papel de la selección del material estimular en el reconocimiento de palabras.

Palabras clave: acceso-léxico, tiempo-reacción, análisis-datos.

Sería una labor ingente resumir en unas páginas —y de manera mínimamente objetiva— todo un campo de conocimiento como es el de la psicología de la lectura, en sus muy diversas vertientes. Además de diferentes textos publicados recientemente en España sobre psicología del lenguaje (v.g., Carreiras, 1997), debemos mencionar varias compilaciones que recogen trabajos recientes de psicología del lenguaje en castellano (Carreiras, García-Albea y Sebastián, 1996; de Vega y Cuetos, 1999). Nuestro interés en este trabajo es mucho más modesto y se ciñe, en primer lugar, a un análisis crítico de las técnicas ideadas para estudiar los procesos de reconocimiento visual de palabras, unas técnicas que, en muchos casos, son compartidas por

* Esta investigación fue subvencionada por una beca de la Dirección General de Investigación Científica y Técnica (PB/97-1379), así como por una beca pre-doctoral del Ministerio de Educación y Cultura para Eva Rosa. La correspondencia sobre este trabajo deberá ser enviada a Manuel Perea. Departament de Metodologia. Facultat de Psicologia. Av. Blasco Ibáñez, 21. 46010-València (Spain). (e-mail: mperea@uv.es)

otras áreas de la psicología cognitiva (v.g., en el campo de la memoria). Y es que, si una buena teoría es importante para el avance científico, no lo son menos unas técnicas adecuadas para poner a prueba dicha teoría. En definitiva ambos, teoría y método, se apoyan la una en el otro y viceversa: los métodos se desarrollan en la dirección que marcan los supuestos de la teoría a la que sirven, mientras que las teorías se verifican sólo en la medida en que lo permiten los métodos empleados para ello. Por otro lado, se tratarán desde un punto de vista crítico los procedimientos de análisis de datos en psicología del lenguaje (que es un tema generalizable a otros ámbitos de la psicología cognitiva) y en el marco general del contraste de hipótesis. Como se expondrá más adelante, los procedimientos de análisis que suelen realizar los experimentadores distan mucho de ser los óptimos.

1. Técnicas de estudio en el reconocimiento visual de palabras.

En las cuatro últimas décadas, el reconocimiento visual de palabras ha sido el área más explorada dentro de la psicología del lenguaje, en comparación con otros temas de mayor interés "ecológico", como podrían ser los aspectos del propio discurso o los análisis de tipo sintáctico y semántico. Una razón para ello es que el reconocimiento visual de palabras es el proceso básico a partir del cual surgen el resto de procesos de lectura. Un adecuado conocimiento de los procesos básicos de la identificación de las palabras es, cuanto menos, una buena base para examinar de manera rigurosa el estudio de procesos que exigen un alto nivel de procesamiento. Además, el conocimiento de los procesos subyacentes al reconocimiento visual de palabras es especialmente importante en la evaluación de los diferentes trastornos de la lectura. De esta manera, diversos modelos teóricos pueden identificar con cierto éxito los estadios del reconocimiento de palabras que pueden estar dañados en los diferentes tipos de dislexia adquirida (v.g., Patterson, 1981). Tal beneficio se puede extender a los programas de rehabilitación, dado que la terapia puede dirigirse directamente a la reconstrucción (o sustitución) de los supuestos déficits del componente que se encuentre lesionado en el sistema de lectura (v.g., Hang, 1996; de Partz, 1986; Plaut, 1996). No podemos ignorar, por otro lado, las ventajas del estudio visual de palabras que derivan de cuestiones más pragmáticas como es la relativa sencillez de las técnicas empleadas. No cabe duda de que trabajar con unidades lingüísticas mayores (oraciones, textos...), o con estímulos verbales auditivos requiere unas técnicas de estudio más sofisticadas.

En definitiva, la especificación de los procesos subyacentes al acceso al léxico proporciona claves de lo que posiblemente es la más clara demostración de las capacidades perceptuales y cognitivas del sistema de procesamiento de información del ser humano. Desgraciadamente, pese a los indudables avances en los últimos 20 años, los aspectos subyacentes a las tareas experimentales distan mucho de estar bien definidos, lo que impide, en ocasiones, tener una adecuada visión de la arquitectura funcional del sistema bajo estudio. Es decir, el tipo de procesamiento que requieren las diversas tareas puede ser relativamente diferente.

1.1 Técnicas taquistoscópicas

Las primeras técnicas experimentales para el reconocimiento de palabras, empleadas predominantemente hacia los años 50 y 60 fueron las tareas de identificación de umbrales, con instrumentos como el taquistoscopio, mediante el cual se presentaba muy brevemente un estímulo que el participante había de identificar. La variable dependiente medida en estas tareas era la precisión de las respuestas bajo una duración estimular constante o bien era el umbral en el que la palabra era identificada. Este tipo de técnicas de identificación taquistoscópica cayó en desuso debido a que ante la presentación breve y/o degradada de los estímulos, los participantes podrían adivinar la palabra a partir de la información sensorial parcial más que propiamente percibirla (para una revisión de los primeros trabajos sobre el tema, véanse Broadbent, 1967; Catlin, 1969). De esta manera, aunque la adivinación involucra la existencia de un acceso al léxico, el proceso quizás no se produce a partir de un flujo de información basado en las características sensoriales del estímulo (Taft, 1991). Los participantes podrían generar un conjunto de posibles palabras candidatas a partir de la información visual parcial y seleccionar cuál es el candidato más probable (v.g. en función de su frecuencia de uso). Por tanto, aunque se pudiera suponer que los procesos señalados anteriormente ocurran también bajo condiciones sensoriales ideales, parece más razonable emplear tareas en las que el material se presente claramente visible, como es el caso de las técnicas presentadas en los siguientes apartados.

1.2 Técnicas cronométricas

Las técnicas o tareas experimentales para el acceso a los procesos mentales que se hallan más en boga desde los años 70 son las denominadas técnicas cronométricas, en las que los estímulos suelen presentarse sin degradar y con duración ilimitada hasta que el participante efectúa la respuesta. En este caso, la variable dependiente de mayor interés suele ser el tiempo de reacción (tiempo transcurrido desde que se presenta el estímulo-test hasta que el participante da una respuesta), aunque también se mide la precisión de las respuestas. El objetivo de estas técnicas consiste en medir el curso temporal de los procesos mentales. A la base de estos procedimientos cronométricos hay una serie de supuestos básicos (Forster, 1979):

1) El procesamiento del lenguaje puede ser analizado como una secuencia de operaciones, cada una de las cuales transforma una representación mental de un estímulo lingüístico en una representación mental de diferente tipo.

2) En una misma tarea experimental es posible que varios sistemas de procesamiento diferentes puedan tener suficiente poder como para proporcionar la información necesaria para la decisión. Dependiendo de las circunstancias, la decisión del participante puede estar basada en el resultado de cualquiera de estos sistemas.

3) La decisión estará basada en el resultado del sistema que finalice antes. Es el llamado supuesto de los “modelos de carreras” (*race models*).

No obstante, el procesador más rápido no siempre controla la decisión, dado que depende de al menos dos factores adicionales: a) la accesibilidad de los resultados al sistema de decisión y b) el tiempo requerido por el elemento decisor para inferir la respuesta correcta al ítem presentado.

Dentro de las técnicas cronométricas podemos señalar tres grandes grupos de tareas:

1) Tareas de clasificación. Se presenta a los participantes cierto estímulo lingüístico que debe ser clasificado de acuerdo con algún criterio previo tan rápidamente como sea posible, procurando no equivocarse. Tal criterio sirve para que la clasificación no pueda efectuarse hasta que el participante haya completado cierto procesamiento lingüístico del estímulo.

La tarea de clasificación más comúnmente utilizada es la tarea de *decisión léxica*, en la que a los participantes se les presenta una secuencia de letras, en una pantalla de ordenador, con respecto a la cual deben decidir si forma o no una palabra de un idioma determinado. En el caso de que la secuencia fuera una palabra (v.g. CASA), los participantes deberían pulsar una tecla señalada, mientras que si no lo es (es decir, si es una pseudopalabra, v.g., DISA), los participantes han de pulsar otra tecla. La tarea de decisión léxica se desarrolló a principios de los años 70 por Rubinstein y colaboradores (Rubinstein, Garfield y Millikan, 1970; Rubinstein, Lewis y Millikan, 1971a, 1971b; pero véase también Landauer y Freedman, 1968) y desde entonces ha sido la tarea que ha suscitado mayor atención, aunque también críticas (véase Balota y Chumbley, 1984, 1990 frente a Monsell, Doyle y Haggard, 1989 y Monsell, 1990). Una alternativa a esta tarea es la *decisión léxica "simple"* (*go/no-go*), en la que sólo se ha de responder cuando la secuencia de letras sea una palabra, pero no se precisa dar una respuesta a las pseudopalabras (v.g., Gordon, 1983; Gordon y Caramazza, 1992; Hino y Lupker, 1998; Perea, Rosa y Gómez, 1999). Esta última tarea tiene la ventaja de que los tiempos de reacción son más breves que en la decisión léxica estándar y que las tasas de errores son también menores.

Otra tarea de clasificación que se ha empleado con relativa asiduidad es la tarea de *verificación semántica*. En este caso, se pregunta a los participantes si la palabra que aparece en la pantalla pertenece a cierta categoría (v.g., ¿es la palabra un animal?). En caso de que la palabra efectivamente sea un elemento de tal categoría (v.g. águila), los participantes pulsar la tecla correspondiente a "SÍ", mientras que si no lo es (v.g., tinaja) los participantes pulsar la tecla correspondiente a "NO".

Naturalmente, hay muchas otras posibilidades, como es la *doble decisión léxica* (v.g., Fischler, 1977), en la que los participantes deben indicar si dos secuencias de letras son ambas palabra o no, o la *decisión de género* en la que los participantes deben indicar si una palabra es de género masculino o femenino. Una combinación de ambas tareas consistiría en

indicar si dos palabras concuerdan en género (v.g. el par "LECHE LOCO"), independientemente del significado (v.g., Gollan y Frost, 1999).

2) Tareas de pronunciación. En este caso, el participante debe pronunciar tan rápido como sea posible el estímulo lingüístico que se les presenta en una pantalla de ordenador, el cual está conectado a una llave vocálica. Habitualmente, los estímulos presentados suelen ser únicamente palabras (*tarea de pronunciación estándar*), pero también pueden presentarse pseudopalabras junto con las palabras (*tarea de pronunciación mixta*, v.g., Hudson y Bergman, 1985). Una variante de las tareas de pronunciación es la tarea de *pronunciación demorada* (*delayed naming*), en la que los participantes han de pronunciar el estímulo, no en el momento en que lo han identificado, sino tras oír cierta señal (por ejemplo, tras aproximadamente 1 segundo de la aparición del estímulo). El objetivo de esta tarea consiste en analizar posibles efectos en la pronunciación de los estímulos más que en el acceso al léxico.

3) Tareas de identificación. El participante debe indicar, pulsando una tecla, el momento en que es capaz de identificar un estímulo, el cual se presenta inicialmente degradado y progresivamente va apareciendo con mayor claridad a medida que la ratio señal/ruido va disminuyendo. La tarea de esta tipo más utilizada en los últimos años es la tarea de *desenmascaramiento progresivo* (Grainger y Seguí, 1990). La ventaja de este procedimiento respecto a los procedimientos "taquistoscópicos" de identificación es que se puede medir el tiempo de reacción. Además, las tasas de errores con esta tarea son muy bajas, lo que parece indicar que no hay evidencia clara de adivinación por parte de los participantes.

Naturalmente, la clasificación anterior no deja de ser una mera aproximación, ya que hay una gran variedad de técnicas "mixtas", tal como por ejemplo la tarea de pronunciación de un estímulo lingüístico sólo en el caso de que el estímulo sea palabra, tarea que tiene un componente de clasificación (decidir si el estímulo es o no una palabra) así como un componente de pronunciación (v.g., véase Hino y Lupker, 1998).

1.3 Métodos de cronometraje de lectura

Este tipo de métodos se basa en el cronometraje del tiempo que dedica el participante a la lectura de material estimular, presentado de distintas formas según la técnica concreta. Este cronometraje se considera una medida indirecta del curso temporal del procesamiento interno. La interpretación de los tiempos de lectura se basa en dos hipótesis (Haberlandt, 1994): la hipótesis de la inmediatez, y la hipótesis de la simultaneidad ojo-mente. La hipótesis de la inmediatez supone que el participante procesa unidades estimulares, usualmente palabras, justo después de ser abordadas, en lugar de esperar al final de una oración o unidad mayor para llevar a cabo el procesamiento. La hipótesis de la simultaneidad ojo-mente afirma que no hay

demora temporal entre el momento en que una palabra es focalizada por el ojo, y el momento en que esta misma palabra está siendo procesada.

El método de cronometraje de lectura más utilizado es el de seguimiento ocular (aunque hay variedades, como el método de ventana móvil, en el que el texto se presenta de manera parcial y no se recogen los movimientos oculares¹; véanse Danks, 1986; Haberlandt y Graesser, 1990). Este tipo de métodos ha sido utilizado principalmente para el estudio de la lectura normal de textos, en silencio (véase Rayner, 1998, para una revisión). El método de seguimiento ocular consiste en presentar al participante experimental un texto en la pantalla de un ordenador para que éste lo lea. Mientras tanto, los movimientos oculares del participante son captados mediante la proyección de un haz de luz invisible sobre la córnea de uno o ambos ojos (aunque también existen otros procedimientos de seguimiento ocular). El análisis de estos registros muestra algunos aspectos interesantes acerca de los movimientos oculares durante la lectura. Lo primero que puede observarse es que los dos ojos se mueven de forma sincrónica a través del texto (Rayner 1978). Pero este movimiento no es continuo sino periódico: a un periodo de movimiento sigue otro de pausa y así sucesivamente. Los periodos de reposo tienen una duración media de 250 ms. (con un rango de oscilación de 150 a 500 ms.) y se denominan fijaciones. Los movimientos, denominados sacádicos, son rápidos y de tipo balístico, lo que quiere decir que una vez iniciados no pueden ser modificados. La duración media de estos movimientos es de 20-35 ms. aunque varía dependiendo de la distancia recorrida. La mayoría de los movimientos sacádicos son de izquierda a derecha (excepto en idiomas como el árabe o el hebreo cuya escritura es de derecha a izquierda) aunque aproximadamente del 10 al 15 % de los movimientos son regresiones, es decir, movimientos hacia atrás en el texto (Just y Carpenter, 1980; McConkie, Hogaboam, Wolverton, Zola y Lucas, 1979; Rayner, Sereno, Morris, Schmauder y Clifton, 1989). El participante extrae la información visual del texto durante las fijaciones, nunca durante los movimientos. Además, algunas palabras reciben más de una fijación, mientras que otras no reciben ninguna.

El empleo de técnicas de seguimiento ocular tiene la ventaja de que se puede precisar el curso temporal del efecto de la variable independiente (véase Perea y Pollatsek, 1998). Si el efecto incide sobre el porcentaje de veces en que se "salta" la palabra-test o sobre la duración de la primera fijación de la

¹ Aunque los métodos de ventana móvil son menos ecológicos que los de seguimiento ocular han demostrado su eficacia para detectar diferentes efectos de interés en el reconocimiento de palabras (v.g., efectos de frecuencia y de longitud de las palabras, etc.). Sin embargo, los diversos estudios comparativos entre los estudios de ventana móvil y de seguimiento ocular muestran resultados muy diversos, desde aquellos que estiman altas correlaciones de los tiempos de lectura obtenidos con ambos métodos (Just, Carpenter y Woolley, 1982) hasta los que afirman que los tiempos de lectura estimados con el método de ventana móvil son un 80% mayores que los estimados con métodos de seguimiento ocular (v.g., Clifton, Speer y Abney, 1991).

palabra-test, se podría indicar que el efecto es "temprano". En cambio, si sólo se observa en la palabra siguiente a la palabra-test o en el porcentaje de regresiones hacia ésta, se podría concluir que el efecto es "tardío", posiblemente de integración contextual. De hecho, es habitual en los estudios sobre movimientos oculares, que se recojan un buen número de variables dependientes (v.g., la duración de la primera fijación sobre la palabra-test, la duración de la mirada sobre la palabra-test antes de ir a la palabra siguiente, la duración total de las fijaciones sobre la palabra-test, el porcentaje de ocasiones en que se "salta" la palabra-test, el porcentaje de ocasiones en que se "vuelve" a la palabra-test, entre otras), con objeto de captar el curso temporal del procesamiento lingüístico.

1.4 Técnicas neurofisiológicas

En la actualidad empieza a hacerse cada vez más frecuente el uso de una nueva generación de técnicas, las neurofisiológicas, en el marco de la aproximación de la neurociencia cognitiva. Y es que es indudable que una adecuada comprensión de los procesos cognitivos necesita cierta comprensión del "hardware" en que tienen lugar estos procesos. Este tipo de técnicas intenta evaluar los procesos mentales mediante correlatos de la actividad cerebral. Uno de los procedimientos más utilizado es el de potenciales evocados (v.g., Kutas y Hillyard, 1980; Rugg y Nagy, 1987; Van Petten y Kutas, 1987). Los potenciales evocados son cambios eléctricos iniciados en un organismo cuando se le presenta un estímulo, en este caso un estímulo lingüístico visual o auditivo. Estos cambios son registrados mediante macroelectrodos situados cerca de las vías sensoriales que conducen hasta el cerebro la actividad eléctrica desencadenada por el estímulo. Estos macroelectrodos llevan a cabo un registro relativamente grueso de la señal eléctrica, que llega además con bastantes "ruidos"; por lo que es un método bastante impreciso para determinar la procedencia exacta de la señal. Algo que aparece relativamente claro en estos registros son una serie de picos que presenta la señal eléctrica en determinados momentos posteriores a la presentación del estímulo, que aproximadamente siempre son los mismos. En concreto, se han identificado picos que parecen relacionarse con distintos procesos mentales. Por ejemplo, a los 300-400 ms. se produce un pico positivo (el llamado P300) que podría tener como correlato la decisión del participante acerca de qué respuesta seleccionar. Por su parte, el pico negativo que ocurre hacia los 400 ms (el componente N400) se suele asociar al reprocesamiento de información semánticamente anómala (Kutas y Hillyard, 1980).

Además del uso de potenciales evocados, ha habido una serie de estudios que han empleado técnicas de imágenes de la corteza cerebral, como las obtenidas mediante resonancia magnética (*functional magnetic resonance imaging*, fMRI; v.g., Pugh et al., 1997) o emisión tomográfica de positrones (*positron emission tomographic*, PET; v.g., Demonet, Price, Wise y Frackowiak, 1994; Petersen et al., 1989). Aunque una descripción detallada de las mismas nos llevaría más allá de los objetivos de este trabajo, con la

tecnología de imágenes cerebrales ha sido posible aislar las regiones de la corteza cerebral que están asociadas a los componentes de procesamiento en diferentes tareas cognitivas. En este ámbito, podemos citar estudios, como el de Pugh et al. (1997) que tratan de enlazar la tradición cronométrica con medidas neurofisiológicas. Para ello compararon los resultados de los mismos participantes en una tarea de decisión léxica y en una tarea en la que recogieron imágenes cerebrales mediante fMRI. El objetivo fue derivar modelos cognitivos del proceso de reconocimiento visual de palabras a partir de los resultados convergentes de tareas cronométricas y medidas neurofisiológicas.

1.5 ¿Cuál es el mejor procedimiento?

Como ha quedado expuesto, desde la década de los 50 hasta la actualidad han sido muchas las técnicas que han ido sumándose al potencial metodológico de experimentación en psicología cognitiva de manera general, y del estudio del procesamiento lingüístico en particular. Sin embargo, a pesar del tremendo avance tecnológico de las dos últimas décadas, ningún método ha escapado a la crítica; todos parecen poseer características inherentes que les impide ser imparciales a la hora de juzgar una determinada teoría. Este problema ha llevado a los investigadores a un mayor esfuerzo para lograr controlar los posibles efectos derivados de las técnicas específicas. Sin embargo, los autores no se ponen de acuerdo en cuál sea la más adecuada o, en otros términos, la menos contaminada por posibles procesos idiosincrásicos (v.g., procesos de decisión, pronunciación, adivinación, etc.). Por todo ello, en la actualidad cabe considerar a todas las tareas experimentales como "interinas" o provisionales. Veamos algunos problemas de cada una de ellas:

1) Los métodos taquistoscópicos y/o de identificación, como se ha señalado previamente, tienen un componente de adivinación que no puede ser despreciado dado que el material se presenta de forma breve y/o enmascaradamente (Monsell, 1991), incluso cuando el tiempo de reacción es la variable dependiente. Es más, dependiendo del tipo de tarea de identificación, los efectos de una misma variable pueden variar de la facilitación a la inhibición (por ejemplo, en el trabajo de Snodgrass y Mintzer, 1993), lo que implica la existencia de ciertos efectos específicos de la tarea que no siempre son fáciles de averiguar. Por ello, los experimentos que emplean estos procedimientos deberían ir acompañados de experimentos con alguna otra tarea que pueda dotarlos de validez convergente y su uso, en la actualidad, es bastante limitado.

2) De los métodos de clasificación se ha cuestionado, en primer lugar, su naturaleza intrusiva. Idealmente, cualquier método debería valorar los procesos que pretende estudiar en su estado normal (lectura). Sin embargo, muchas veces los propios métodos constituyen una intrusión en dichos procesos provocando su interrupción y la aparición de procesos adicionales que no ocurren en la lectura normal de textos. En los métodos de clasificación se somete al participante a una situación de test en la que debe responder a un

determinado estímulo, lo cual supone un aumento en la demanda de recursos, y una posible interferencia en los procesos de acceso al léxico. Asimismo, la situación de test puede introducir procesos adicionales distintos de los de acceso al léxico como, por ejemplo, estrategias utilizadas por el participante para aumentar la precisión de sus respuestas en la tarea de decisión léxica (Balota y Chumbley, 1984).

Por lo que respecta a la tarea de clasificación más comúnmente empleada (decisión léxica estándar), pensamos que una alternativa más adecuada sería la decisión léxica "simple" (aquella en la que el participante sólo ha de responder "sí" a las palabras, pero no ha de efectuar respuestas a las pseudopalabras). En ambas tareas, los participantes han de acceder a la representación en su propio léxico del estímulo-test y efectuar la respuesta; sin embargo, al utilizar la tarea de decisión léxica simple se evita un proceso de decisión de respuesta, que está presente en la tarea estándar: la decisión de qué tecla corresponde a la respuesta adecuada. Como los procesos de respuesta son más sencillos en la decisión léxica "simple", se produce un claro descenso en los errores y en los tiempos de reacción (Gordon, 1983; Perea et al., 1999). Respecto a la tarea de verificación semántica (v.g., ¿es la palabra un animal?), un grave problema es que los estímulos-test suelen corresponder a la respuesta "no", y si bien existe cierta evidencia de cómo los participantes ejecutan la respuesta "sí" (acceso a la información lingüística de la palabra) no lo está tanto el estudio de cómo los participantes deciden contestar "no" (es decir, la respuesta por defecto) en una tarea de verificación semántica. Además, cabe la posibilidad de que los procesos de acceso al léxico en estas tareas de verificación no sigan los caminos usuales que dan lugar al reconocimiento de palabras.

3) El problema básico en las tareas de pronunciación es que, en muchos idiomas (entre los que se encuentra el castellano) puede realizarse sin necesidad de que haya un procesamiento léxico. Cualquier persona es capaz de pronunciar una pseudopalabra castellana correctamente, o una palabra que nos es desconocida, y por tanto no está almacenada en nuestro léxico interno. Para ello, simplemente seguiríamos las normas de correspondencia entre grafema y fonema (Baluch y Besner, 1991; Forster, 1981; Halle, 1990; Tabossi y Laghi, 1992). Esto supone que cuando medimos el tiempo de reacción en la tarea de pronunciación no podemos estar seguros de que corresponda al tiempo de acceso al léxico. En este sentido, algunos resultados experimentales sugieren la existencia de un código fonológico pre-léxico que actúa muy rápidamente durante la lectura (Perfetti, Bell y Delaney, 1988). No obstante, parece que el procesamiento de palabras es fundamentalmente léxico en una lengua como el castellano, como se deriva del hecho de encontrar facilitación de tipo asociativo/semántico en pruebas de pronunciación (v.g., Perea y Gotor, 1997; Sebastián-Gallés, 1991). No obstante, estos efectos son usualmente de menor tamaño que con la tarea de decisión léxica. Un segundo problema de las tareas de pronunciación es que la pronunciación de una palabra relativamente larga puede comenzar antes de que su procesamiento léxico haya sido completado. Con tener el segmento (o mejor, la sílaba) inicial, el participante puede iniciar su respuesta.

4) Los métodos de seguimiento ocular, que aparentemente son la alternativa más deseable a la hora de estudiar la lectura, por ser los menos intrusivos y los más cercanos a los procesos de lectura, tampoco se hallan exentos de críticas. La primera es la que pone en duda la hipótesis de la inmediatez citada anteriormente. Como ya se indicó, esta hipótesis supone que los participantes se centran cada vez en una unidad de texto, usualmente palabras, y la procesan inmediatamente sin esperar a un momento posterior cuando se haya completado una unidad mayor, como por ejemplo una oración. Sin embargo, es habitual encontrar lo que se denominan efectos de arrastre (*spillover effects*): en ocasiones los lectores continúan procesando una palabra (o unidad mayor) previa, al tiempo que ya están inspeccionando una nueva unidad de texto (Haberlandt y Bingham, 1978; Perea y Pollatsek, 1998; Rayner et al., 1989), o pueden anticiparse previendo el segmento de texto subsiguiente al que están leyendo (Rayner et al., 1989). No obstante, este argumento ya se encuentra incluido en el más reciente modelo de lectura y movimientos oculares del grupo de *Massachusetts (E-Z Reader model*; Reichle, Pollatsek, Fisher y Rayner, 1998). Una segunda crítica a la experimentación en que se registran los movimientos oculares es la que se refiere a la plausibilidad de las oraciones. Cabe la posibilidad de que las oraciones difieran en plausibilidad más como consecuencia de la propia manipulación. Por ello, es aconsejable que se obtengan datos de lo “naturales” que son las sentencias antes de utilizarlas como material experimental. Una tercera crítica es que los procedimientos de seguimiento ocular son poco versátiles en el sentido de que, caso de disponer de muestras relativamente escasas de ítems por condición experimental, sería francamente difícil lograr oraciones igualmente plausibles en todas las condiciones. Si tuviéramos “gnomo” y “mamut” como palabras-test en dos condiciones diferentes (en función de, por ejemplo, la frecuencia de uso de la primera sílaba; Carreiras, Álvarez y de Vega, 1993; Perea y Carreiras, 1998) se podría construir una frase como esta “Pedro lloró la muerte del último gnomo/mamut en la película de ayer”. Evidentemente, esta estrategia es difícil de llevar a cabo cuando tengamos tres o más condiciones experimentales o cuando el número de ítems por condición sea pequeño.

5) Las tareas que emplean técnicas neurofisiológicas (especialmente los potenciales evocados) suelen ser puestas en duda debido a que, al menos en estos momentos, no se conoce muy bien lo que están midiendo (Rayner y Pollatsek, 1989). No obstante, no cabe duda que los resultados de estos estudios pueden poner restricciones respecto a los tipos de procesos cognitivos que son biológicamente plausibles. En todo caso, las técnicas que emplean imágenes cerebrales (v.g., PET, fMRI) no poseen una resolución temporal alta, a diferencia de los potenciales evocados. Por su parte, los potenciales evocados, a pesar de su excelente resolución temporal, normalmente no permiten aislar con precisión las regiones concretas del cerebro que producen un determinado efecto. De esta manera, el empleo de los potenciales evocados podría complementar los estudios de técnicas que, como el fMRI, tienen una alta resolución espacial, pero no temporal. Desgraciadamente, en la actualidad resulta difícil para muchos laboratorios

disponer de la infraestructura necesaria para efectuar este tipo de experimentos (con la salvedad de la técnica de potenciales evocados, de amplia utilización), lo que puede dificultar la replicación de los hallazgos a través de diferentes laboratorios.

En definitiva, cada método posee sus propias ventajas e inconvenientes, de manera que no hay ninguno perfecto. Todos ellos captan los procesos centrales de la identificación de palabras, junto con otra serie de procesos específicos del propio método. Esto es lo que Grainger y Jacobs (1996) han denominado "solapamiento funcional". Un ejemplo claro de este solapamiento se puede observar en el hecho de que cada método hace una estimación distinta del tiempo de reconocimiento de una palabra escrita (ver Rayner y Pollatsek, 1989; Schilling, Rayner y Chumbley, 1998). Ante este panorama, suele recomendarse la utilización de varios métodos para poner a prueba un mismo supuesto teórico, y de este modo dar mayor validez a los resultados experimentales (v.g., Carpenter, 1984; Haberlandt, 1994; Rayner, Flores d'Arcais y Balota, 1990; Rayner y Pollatsek, 1989; Singer, 1990).

De esta manera, si los resultados son similares a través de tareas cabría pensar que el efecto bajo estudio es típicamente léxico y no es debido a cierta característica de una sola tarea experimental. No obstante, tal argumento tiene sus problemas cuando sólo se ciñe a algunas tareas y se efectúa de manera ciega. Veamos un ejemplo: los efectos del "número de palabras ortográficamente similares" a la palabra-test son facilitadores tanto en decisión léxica como en pronunciación, lo cual fue utilizado por Andrews (1989) para afirmar que dichos efectos eran "reales". Sin embargo, Grainger y Jacobs (1996) han mostrado que tales efectos podrían ser debidos a procesos específicos de cada una de estas dos tareas. De hecho, en una reciente serie de experimentos, Pollatsek, Perea y Binder (1999) encontraron efectos inhibidores del número de palabras similares en una prueba de lectura en que se medían los movimientos oculares (la prueba que se supone que mejor refleja los procesos de lectura), que iba acompañada de efectos facilitadores en una tarea de decisión léxica.

En todo caso, se debe tener en cuenta cuáles son los procesos subyacentes a las tareas, para lo cual se precisa una arquitectura funcional con los pasos requeridos para efectuar la tarea, y cuáles son los fenómenos de interés. Así, por ejemplo, en caso de que quisiéramos comprobar si se produce una codificación fonológica de tipo automático en las palabras, sería interesante comparar los resultados de la tarea de decisión léxica y la tarea de pronunciación, dado que si tal efecto sólo ocurriera en la tarea de pronunciación, el efecto podría deberse únicamente a las características inherentes de esta tarea. De este modo, sería difícil separar un estadio de fonología de "entrada", respecto de un estadio fonológico de tipo articulatorio.

Para terminar, tan importante como la elección del método es la selección del material estimular y la planificación de controles experimentales adecuados para neutralizar los procesos específicos de la tarea que se esté

utilizando, así como la elección del modo de análisis de los datos experimentales, pero esto último forma parte de la sección siguiente.

2. Análisis de los datos experimentales.

En esta sección nos ocuparemos de tres temas referidos al análisis estadístico de los datos experimentales. El primero, y más general, se refiere a la propia lógica de las pruebas de contraste de hipótesis. El segundo se ocupa del análisis de la variable dependiente más común en psicología cognitiva (y en psicología de la lectura), como es el tiempo de reacción. El tercero se refiere a las posibilidades de generalizar nuestros resultados, no sólo a otros participantes, sino también a otros elementos lingüísticos.

2.1 Críticas a la lógica de contraste de hipótesis

Típicamente, en los experimentos se obtiene (para cada efecto o contraste) una razón F a partir de los datos y, de acuerdo con el valor crítico que adoptemos (tal que la probabilidad de rechazar la hipótesis nula cuando es verdadera sea de 0'05), se decide si mantenemos la hipótesis nula o si, por el contrario, indicamos que ha habido un efecto de la variable independiente. El supuesto que está a la base de esta decisión es que si el valor del estadístico de contraste obtenido con nuestros datos supera el valor crítico, muy posiblemente haya habido una variabilidad debida al factor manipulado; en caso contrario, mantenemos que no hay efecto. Esta lógica de contraste de hipótesis que parece, a primera vista, simple y correcta no se halla exenta de problemas. Desde los años 30 (Berkson, 1938), el número de trabajos que han criticado la lógica del contraste de hipótesis ha ido en aumento (véase la reciente compilación de Harlow, Mulaik y Steiger, 1997). Sin embargo, tales críticas han tenido poca influencia, lo que puede comprobarse fácilmente si se acude a las revistas de psicología experimental. Incluso una revista como *Memory and Cognition*, cuyo director desde 1993 hasta 1996, Loftus, es un claro opositor de la lógica del contraste de hipótesis sirve de muestra para indicar la relativa poca influencia de estas críticas.

Y es que sigue habiendo una enorme resistencia por parte de los investigadores —que no ya de los estadísticos— respecto a dejar de usar el contraste de hipótesis. Como dice Oakes (1986): "Muchos investigadores tienen un cariño desmesurado a las pruebas de significación a pesar de los argumentos tan contundentes que se han presentado en contra de las mismas, en los capítulos 1-3 recogeré tales argumentos...en un intento de matar la bestia —pero sospecho que el cuerpo sin cabeza continuará azotando las páginas de las revistas científicas en los años venideros". En el mismo contexto, Harris (1991) señalaba "dada la vehemencia y repetición de tales ataques es sorprendente que el dragón no muera". Seguidamente, se presentan varias de las críticas más comunes a la lógica del contraste de hipótesis.

2.1.1 El contraste de hipótesis no nos dice lo que queremos saber

Las pruebas de contraste de hipótesis y la inferencia científica tratan cuestiones diferentes. En la inferencia científica lo que queremos saber es cuál es la probabilidad de que la hipótesis nula (H_0) sea cierta dados ciertos datos (D); es decir, $P(H_0/D)$. Sin embargo, lo que las pruebas de contraste nos proporcionan es la probabilidad de obtener datos como los obtenidos (o más extremos) cuando la hipótesis nula es cierta, $p(D/H_0)$. Lógicamente, la obtención de un $p(D/H_0)$ pequeña no implica necesariamente que $P(H_0/D)$ sea también pequeña.

El razonamiento de los investigadores es un "pseudo" *modus tollens*, que ha sido descrito por Falk y Greenbaum (1995) como una "ilusión de una prueba probabilística por contradicción". Siguiendo a uno de los primeros críticos de la lógica del contraste de hipótesis, Berkson (1942) decía que "en primer lugar el argumento resulta ser ilógico. Considerémoslo en una forma simbólica. Si se dice 'Si A es cierto, B ocurrirá raramente; en consecuencia, si se encuentra B, podemos desacreditar A'. No hay ninguna garantía lógica para considerar que un suceso que ocurra, aunque sea infrecuente, pueda desacreditar la hipótesis" (p. 326).

Además, podríamos destacar cierto número de afirmaciones relacionadas con el contraste de hipótesis que son falsas: 1) que una p muy baja da una idea de la magnitud del efecto o de la relación entre variables; 2) que el valor de p es la probabilidad de que la hipótesis nula sea correcta; 3) que el complemento de p (es decir, $1-p$) es la probabilidad de que se obtenga un resultado significativo en una replicación del estudio. Evidentemente, el valor de p únicamente muestra la probabilidad de que los resultados sean al menos tan extremos como los obtenidos con esos datos, suponiendo que la hipótesis nula sea cierta. El valor de p depende de diversos factores. En el caso de las razones F , la media cuadrática del numerador es elevada respecto a la media cuadrática del denominador cuando el tamaño del efecto es grande, el número de valores por condición es elevado, o cuando ambos valores son elevados.

2.1.2 El contraste de hipótesis es una prueba trivial

La premisa básica de la que partimos es que la hipótesis nula es prácticamente siempre falsa en psicología —con excepciones como el caso de la percepción extra-sensorial y temas parecidos. Como dijo Cohen (1990, p. 1308), "la hipótesis nula, tomada literalmente (y esta es la única forma de entenderla en el marco de la lógica del contraste de hipótesis) es siempre falsa en el mundo real." De ahí se deduce que cualquier decisión de rechazarla indica simplemente que el diseño experimental tuvo la potencia estadística adecuada para detectar el efecto, el cual podrá ser grande o pequeño, de gran importancia o meramente secundario. Es curioso que la adherencia a los principios del contraste de hipótesis haya dado lugar a una probabilidad de error de tipo I muy baja, mientras que la probabilidad de error de tipo II se

mantiene en valores más altos de los deseables. De hecho podemos leer en fecha tan temprana como 1938 lo siguiente: "si los resultados de la prueba son conocidos, no es una prueba de hecho" (Berkson, 1938, p. 527).

2.1.3 La decisión dicotomizada

Al adoptar un nivel de significación fijo (usualmente 0'05), el investigador pasa de un continuo de certidumbre-incertidumbre a una decisión dicotómica de rechazar o no rechazar la hipótesis nula. Evidentemente, es difícil de ver la diferencia real entre un experimento con una p de 0'049 y otro con una p de 0'051 —salvo que en uno se rechaza la hipótesis nula y en el otro no—. Un problema adicional es que algunos investigadores, de manera equivocada, interpretan el hecho de no rechazar la hipótesis nula como evidencia de que la hipótesis nula es cierta.

Además, las pruebas de contraste de hipótesis están sesgadas contra la hipótesis nula. Un comentario familiar de todos aquellos experimentadores que emplean la lógica del contraste de hipótesis es que este tipo de pruebas sólo permite una posible dirección en los resultados: el rechazo de la hipótesis nula. De esta manera, se oyen frases como "no se puede probar la hipótesis nula". Con esta lógica, se puede rechazar la hipótesis nula e indicar que es errónea. Por el contrario, en caso de no rechazar la hipótesis nula no tenemos certeza de lo que está ocurriendo, lo cual es sin duda influjo de la lógica fisheriana (v.g., Algarabel, Meseguer, Dasí y Ruiz, 1989).

Es decir, el rechazo de la hipótesis nula da apoyo a la hipótesis alternativa, mientras que un "no rechazo" de la hipótesis nula no apoya la hipótesis nula ni rechaza la alternativa. Existen varias razones para tal asimetría (Greenwald, González, Harris y Guthrie, 1996): a) la hipótesis nula es usualmente una hipótesis puntual, mientras que la hipótesis alternativa suele ser un rango de hipótesis; b) usualmente la probabilidad de error de tipo II suele exceder la probabilidad de error tipo I; y c) cabe la posibilidad de que el "no rechazo" de la hipótesis nula haya sido consecuencia de un experimento defectuoso, con poca potencia, o con demasiada variabilidad.

2.1.4 Alternativas al contraste de hipótesis

A nuestro juicio, el tema de la adecuación del contraste de hipótesis a la investigación científica no siempre ha de tomarse en sentido absoluto, sino que dependiendo del campo o ámbito, el énfasis puede ponerse en unos aspectos más que en otros. Todo ello se agrava debido a que el proceso de contrastes de hipótesis en psicología es una mezcla, en muchos casos inconsistente, de la tradición fisheriana con la tradición de Neyman-Pearson (para una revisión, véase MacDonald, 1997). Sin embargo, el hecho de que el contraste de hipótesis sea insuficiente en ciertos casos no implica que no tenga un papel de importancia en el marco científico (Frick, 1996). Evidentemente, lo que queremos es poder generalizar a partir de nuestros datos con el objeto de tratar de construir teorías psicológicas.

Una buena alternativa (o complemento) al contraste de hipótesis es el uso de intervalos de confianza. El cómputo de un estimador puntual de la diferencia entre las medias del grupo A y del grupo B y un intervalo de confianza para tal diferencia no requiere mayor información de la necesaria para una prueba de contraste de hipótesis. Un intervalo de confianza contiene toda la información proporcionada en una prueba de contraste de hipótesis y, además, proporciona un rango de valores dentro de los cuales es muy probable que se encuentren las diferencias "verdaderas". Además, un estimador puntual y un intervalo de confianza emplean la misma unidad de medida de los datos, lo que facilita la interpretación de los resultados. Desgraciadamente, a pesar de la superioridad —no sólo en términos de información— de los intervalos de confianza, no suelen aparecer con frecuencia en las revistas psicológicas. Al contrario, lo que vemos en estas revistas es una lógica ciega de rechazar/no rechazar, que no nos proporciona muchas veces la información que queremos saber, y una particular preocupación en los valores de p .

Por otra parte, no debemos olvidar un criterio como la replicación. La presión por publicar hace que, en ocasiones, los investigadores no repliquen siempre los estudios en que obtienen resultados novedosos. De hecho, todo aquel que ha realizado experimentos habrá observado, tarde o temprano, que resultados que en un primer estudio eran significativos, en un segundo estudio —con características similares— no lo han sido y viceversa. Aunque éste sea un criterio extra-estadístico, no cabe duda que es de una importancia esencial y es, sin duda, la piedra de toque para los investigadores experimentales. Lógicamente, si un hallazgo se replica en diferentes laboratorios quiere decir que el hallazgo es consistente, y en ese caso, la estadística tiene un papel menor. Piénsese que muchos de los más famosos psicólogos de este siglo no se caracterizaron por emplear profusamente la estadística (v.g., Piaget, Skinner, entre otros).

Finalmente, y no por ello menos importante, las investigaciones meta-analíticas pueden ser de suma importancia en este contexto (véase Schmidt, 1996), aunque un tratamiento pormenorizado de este tema nos llevaría más allá de los objetivos de esta revisión. Sin embargo, un problema de este tipo de investigaciones en nuestra área reside en la gran cantidad de aspectos específicos de las manipulaciones experimentales realizadas (v.g., los niveles de la variable manipulada, técnica empleada, instrucciones, etc.), así como las características idiosincrásicas del idioma en que se lleven a cabo los experimentos, lo cual dificulta la generalización de los resultados.

Para acabar este apartado nos remitimos a Abelson (1997), quien señaló cinco criterios para juzgar la calidad de los resultados de una investigación, que correspondían al acrónimo *MAGIC*: 1) Magnitud (o tamaño) del efecto; 2) Articulación del efecto; 3) Generalización del efecto; 4) Interés del efecto; y 5) Credibilidad del efecto. De acuerdo con Abelson, de estos cinco criterios, el valor de p sólo tiene relevancia respecto a la credibilidad (en el caso de un hallazgo nuevo) y a la generalidad (en el caso de la replicación). Asociado con la credibilidad hay determinados factores, tales como la coherencia teórica, la plausibilidad de los procesos psicológicos

postulados y, quizás menos importante, el valor de p obtenido con una prueba de significación.

2.2 Análisis de los tiempos de reacción

En los experimentos habituales en psicología del lenguaje (así como en otras áreas de la psicología cognitiva), se obtienen múltiples observaciones por participante y condición, siendo el tiempo de reacción la variable dependiente por excelencia. El procedimiento que se sigue es promediar los tiempos de reacción de las respuestas correctas para cada una de estas observaciones, de modo que se obtiene una medida del rendimiento del participante en cada condición experimental. El problema reside en que es habitual encontrarse con algunas observaciones atípicas en el conjunto de datos, debidas bien a distracciones de los participantes o a fallos en el instrumental. Indudablemente, en estos casos, la media aritmética de la condición experimental se vería muy afectada por dichas observaciones atípicas, por lo que los investigadores suelen efectuar diversos procedimientos de recorte de datos antes de realizar los promedios. Sin duda, el problema de qué hacer con las puntuaciones atípicas es un tema de gran interés en el ámbito de la estadística aplicada (véase Barnett y Lewis, 1995). Pero hay un segundo problema de importancia en las investigaciones que emplean el tiempo de reacción como variable dependiente. Las distribuciones de los tiempos de reacción no suelen ser simétricas, sino que, como es bien conocido, suelen mostrar cierta asimetría positiva (Luce, 1986).

A modo de póliza de seguros (Anscombe, 1960), los psicólogos experimentales suelen emplear diversos procedimientos de recorte de datos (v.g., medias truncadas: exclusión de datos más allá de ciertos límites como 250-1500 ms; medias restringidas: exclusión de datos que se encuentren más allá de dos desviaciones típicas de la media del participante; medianas,...) o de transformación (v.g., medias armónicas o medias logarítmicas). Esta es, sin duda, una situación poco deseable porque, como señalaron Ulrich y Miller (1994), se podría sospechar que los resultados están afectados por la técnica de recorte/transformación de datos empleada. Es posible que los investigadores pudieran estar cometiendo un error de tipo I o bien un error de tipo II. Esto hace necesario un estudio detallado de los procedimientos más aconsejables de recorte de datos. En este sentido, el trabajo de Ratcliff (1993) mostró que los diversos procedimientos de recorte que efectúan los psicólogos cognitivos son robustos (es decir, las tasas de error de tipo I son muy similares a las tasas nominales) y que los distintos procedimientos de recorte difieren en su potencia. Perea y Algarabel (1999) han generalizado los hallazgos de Ratcliff a un mayor número de procedimientos de recorte de datos, al tiempo que han estudiado la influencia de la asimetría de la distribución de datos. Las conclusiones de estos dos trabajos coinciden en señalar la técnica de medias truncadas, frecuentemente utilizada, como una técnica potente; aunque en el caso de que existiera gran variabilidad entre las

puntuaciones de los distintos participantes, el empleo de medias restringidas sería la alternativa más adecuada.

Ratcliff (1993) señaló una interesante posibilidad a la hora de utilizar medias truncadas. Se trataría de calcular las medias truncadas con diferentes puntos de corte para comprobar si los efectos de la variable independiente son significativos en varios de ellos. Esto no implica, naturalmente, que el investigador vaya efectuando procedimientos de recorte de manera exploratoria para ver si alguno de ellos es significativo. Estaríamos entonces en una situación similar a la que nos encontramos cuando obtenemos un valor crítico para la razón F dependiendo de si el contraste es a priori o a posteriori. Sin duda, la mejor opción en caso de que un resultado ronde la significatividad es la replicación, especialmente cuando el efecto es nuevo. En este sentido, es importante señalar que la probabilidad de replicar un efecto para el cual se haya obtenido una $p=0'0499$ en un primer experimento es de sólo 0'50 (suponiendo que la estimación del efecto sea la correcta), como mostraron Greenwald et al. (1996). De hecho, para que la probabilidad de replicar el hallazgo fuera de 0'80, el valor de probabilidad p del experimento inicial debería ser de 0'005.

Por otra parte, en las últimas décadas, se han propuesto un buen número de estadísticos robustos (Hampel, Ronchetti, Rousseeuw y Stahel, 1986; Huber, 1981) que, sin embargo, no son utilizados en psicología experimental (véase Wilcox, 1998), prefiriéndose el empleo de medias truncadas o medias restringidas a procedimientos que quizás pudieran ser más adecuados. Hasta ahora no ha habido estudios comparativos de la potencia de los estadísticos robustos frente a los estadísticos comúnmente obtenidos por los psicólogos tras el recorte de datos de tiempos de reacción.

Otro aspecto a tener en cuenta es que en las revistas sobre psicología del lenguaje todavía no parece haber una conciencia clara de lo que supone emplear diseños intra-sujetos en los que algún factor tenga más de dos niveles (v.g., Maxwell y Delaney, 1990). Es bien conocido que en tales casos el procedimiento a seguir es, bien efectuar una corrección en los grados de libertad de la distribución F teórica (según el valor de epsilon obtenido en el ANOVA), o bien seguir la alternativa multivariada. De hecho, en una revista de corte experimental como *Psychophysiology* actualmente se exige a los autores que efectúen un ajuste de los grados de libertad en los artículos que empleen diseños con más de dos niveles en alguno de los factores manipulados intra-sujetos.

Independientemente de lo anterior, en la experimentación actual hay una preferencia desmedida por las pruebas de inferencia sobre las medias, sin tener en cuenta otros aspectos de interés como la forma de la distribución de los datos. Por ello, una opción que puede complementar los análisis de los tiempos de reacción medios por condición es efectuar un análisis de la forma de las distribuciones de datos. Desgraciadamente, hace falta un número bastante elevado de observaciones por condición experimental para poder estudiar la forma de la distribución, lo que no suele ser el caso en muchos trabajos en psicología cognitiva (especialmente por impedimentos derivados

de la selección del material estimular). No obstante, en el caso de tener un número escaso de datos por condición experimental (v.g., 15-20 datos) una opción muy deseable es el procedimiento de “vincentización” (véase Heathcote, 1996; Ratcliff, 1979), en el que a partir de cierto número de cuantiles de cada participante (v.g., los 9 deciles) y de los promedios obtenidos a través de los participantes, se puede obtener un histograma que refleje la forma de la distribución de los tiempos de reacción. En concreto, estos son los pasos para efectuar el histograma “vincentizado”: El primer paso consiste en calcular los cuantiles para cada participante y condición (v.g., los 9 deciles). El segundo paso consiste en promediar los cuantiles a través de los participantes, para así obtener los cuantiles de la distribución agrupada. En tercer lugar, se puede construir fácilmente el histograma de la distribución agrupada (histograma “vincentizado”) situando los cuantiles promediados en la abscisa y levantando sobre los cuantiles adyacentes rectángulos que tengan una misma área (cada rectángulo tendrá un área de $1/q$, donde q es el número de cuantiles). A modo de ejemplo, en la figura 1 se muestra un histograma “vincentizado” empleando los nueve deciles en una distribución de tiempos de reacción.

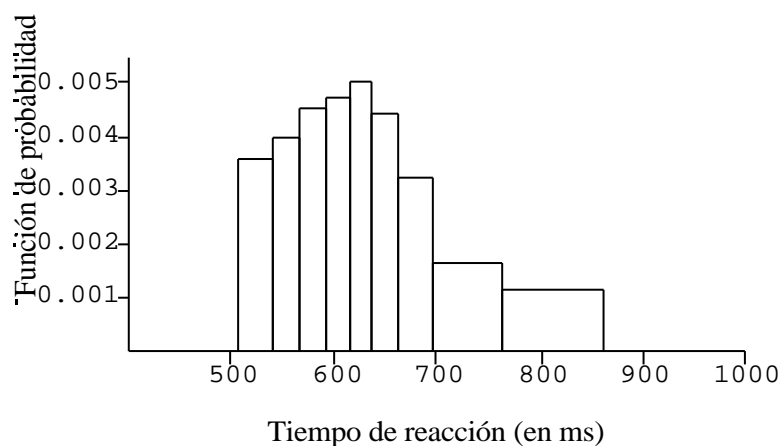


Figura 1. Histograma “vincentizado” empleando deciles en un experimento en que se recogía el tiempo de reacción.

2.3 Sobre la selección y generalización de los análisis a través de ítems.

En los experimentos de psicología de lenguaje el objetivo no es sólo generalizar los efectos obtenidos a otros participantes, sino que también se pretende generalizar los efectos a otros ítems. Si efectuamos únicamente el análisis por participantes tratando los materiales como un efecto fijo, no podemos generalizar los efectos obtenidos a otros ítems. Tras el influyente trabajo de Clark (1973), la gran mayoría de artículos sobre psicología del lenguaje suelen ofrecer tanto un análisis por participantes (donde participantes es la variable aleatoria) como un análisis por ítems, e incluso razones cuasi-F (o también la $F_{mínima}$) que combinan ambos análisis. (No obstante, en la

actualidad es muy poco habitual ver artículos que empleen la razón $F_{\text{mínima}}$.) Hemos de tener en cuenta que si un efecto es significativo en el análisis por participantes pero no lo es en el análisis por ítems, el efecto podría deberse a una o algunas palabras en una de las condiciones que hubieran provocado el efecto en el análisis por participantes, con lo que se limita el alcance de los resultados.

Sin embargo, este es un tema polémico en la actualidad. Como Hino y Lupker (1996, 1998) señalaron, la selección de los ítems en la mayoría de los experimentos de psicología del lenguaje no es aleatoria, sino que los estímulos se seleccionan porque cumplen una serie de criterios (v.g., frecuencia de uso, número de letras, número de sílabas, número de vecinos ortográficos, etc.). Hino y Lupker señalaron que, en tales casos, los análisis por ítems serían inadecuados (véase Keppel, 1982). Aunque el comentario de Hino y Lupker es correcto, no es menos cierto que si el efecto no aparece en el análisis por ítems, esto podría ser indicador de que el efecto se debiera a las características de unos pocos ítems. Además, en los casos en que los resultados son significativos por ítems, nadie argumenta que los análisis por ítems no son necesarios. Pero visto desde otra perspectiva, los comentarios de Hino y Lupker tienen cierto sentido. Pensemos en un experimento en el que se emplean palabras de baja frecuencia bajo dos condiciones (v.g., alta frente a baja concreción). El problema a la hora de efectuar el análisis por ítems residiría en que la media cuadrática de error se calcularía promediando las varianzas estimadas de los tiempos de reacción para los ítems en cada condición. Si hay bastante variabilidad dentro de los ítems de cada condición (algo habitual cuando trabajamos con palabras de baja frecuencia, dado que algunas son más familiares que otras), el denominador de la razón F será bastante grande, con el consiguiente descenso de potencia en el análisis por ítems. De esta manera, cuando el investigador obtiene una razón F significativa por participantes pero no por ítems queda siempre la duda de estar cometiendo bien un error de tipo I o bien un error de tipo II.

En definitiva, hemos de ser muy cautos cuando el efecto sólo aparece en el análisis por participantes, especialmente cuando se emplean palabras de baja frecuencia que son las más susceptibles de ser influidas por terceras variables. Ante esta situación, una alternativa sería efectuar experimentos con pares de ítems que se encuentre emparejados en las variables de control de interés. En estos experimentos el par sería su propio control y las varianzas (suponiendo que los miembros del par hayan sido bien emparejados) serán menores (dado que se tratarían como grupos relacionados de ítems), con la consiguiente ganancia en potencia estadística. De hecho, esta es la lógica que se emplea en experimentos de "priming", en los que en una lista experimental se presenta el par perro-GATO (GATO sería el estímulo-test) y en la otra se presenta el par curva-GATO, actuando cada estímulo-test como su propio control (Pollatsek y Well, 1995), y evitando así el problema de variabilidad a través de los ítems.

Otra opción es efectuar experimentos de "priming", pero aquí podríamos tener un problema adicional. Por ejemplo, si lo que queremos es demostrar que "casa" se activa cuando se presenta el estímulo lingüístico

"CAPA", una técnica de "priming" no es una buena alternativa. El hecho de que los tiempos de reacción del par relacionado casa-CAPA difiera de los del par piso-CAPA no implica necesariamente que "casa" se activa siempre que se presente CAPA. Lo único que indicaría esta diferencia es que el grado de relación entre los miembros del par difiere en ambas condiciones (véase Forster y Taft, 1994).

Finalmente, queda la pregunta de cómo debe efectuarse la selección de los ítems de un experimento. Es claro que si pretendemos generalizar a través de ítems, dichos ítems deben obtenerse de una manera aleatoria (evidentemente, tras los controles adecuados en posibles variables extrañas). Otra razón por la cual es deseable que la selección de ítems sea aleatoria, a partir de un grupo amplio de ítems, consiste en evitar todo posible sesgo del investigador (Forster, 1999; Forster y Shen, 1996). De hecho, Forster (1999) demostró que incluso personas con poca experiencia en experimentación eran capaces de predecir cuál de los miembros de un par de palabras emparejados en frecuencia era el ítem "mejor" en una tarea de decisión léxica.

La cuestión, en todo caso, es si en un determinado grupo de ítems se incluyen todas las palabras que satisfacen todos los criterios puestos por el investigador. Por ejemplo, es bien conocido que si se extraen palabras de baja frecuencia a partir de una base de palabras computerizada (v.g., la base de palabras en castellano de Alameda y Cuetos, 1995), hay algunas que suelen ser desconocidas para el hablante medio del lenguaje y dan lugar a una muy alta tasa de errores. En caso contrario, si se excluyeran, se podría hablar de que el investigador pudiera estar sesgando (siquiera inconscientemente) la muestra de ítems. No obstante, el empleo de la selección totalmente aleatoria de ítems de baja frecuencia puede ser arriesgado, dado que se corre el peligro de que algunas palabras muy poco familiares entren en la muestra final, con el consiguiente aumento en variabilidad, el descenso en potencia y la necesidad de un número mayor de participantes para obtener efectos significativos. En caso de duda, creemos que sería recomendable un pase piloto de aquellas palabras que, habiendo sido seleccionadas previamente al azar, pudieran parecer muy poco familiares a los experimentadores. Si en dicha prueba estos ítems muestran tasas de error altas, pensamos que deberían reemplazarse por nuevos estímulos escogidos aleatoriamente.

3. Conclusiones.

Indudablemente, la lectura es una de las habilidades adquiridas más importantes que ha sido programada biológicamente. Sin embargo, tras varias décadas de intensos estudios sobre el reconocimiento visual de palabras, siguen sin estar resueltas muchas de las cuestiones fundamentales planteadas sobre los procedimientos que rigen el acceso al léxico (véase Andrews, 1997; Forster, 1998). Estos problemas pueden deberse, en parte, tanto a las técnicas utilizadas para medir el acceso al léxico como a la manera en que se han analizado los datos experimentales. Además, aunque en esta revisión no se han analizado, creemos importante recalcar que otros aspectos como las

instrucciones o la propia composición de la lista experimental pueden afectar enormemente a los resultados (v.g., Grainger y Jacobs, 1996).

En la sección primera hemos efectuado una revisión de las técnicas de acceso al léxico en la que concluimos que la alternativa más aconsejable es el empleo de métodos de seguimiento ocular, tanto por su naturaleza poco intrusiva como por su validez ecológica, aunque sería deseable que tales experimentos pudieran replicarse con otras técnicas. En cualquier caso, hay cierto número de fenómenos que son difíciles de estudiar utilizando oraciones como material lingüístico. Por ejemplo, en el estudio de efectos para los que el número de estímulos susceptible de ser utilizado como material experimental es relativamente limitado es muy difícil construir una frase en la que encaje de manera plausible un ítem de cada condición experimental. Además, en experimentos de "priming" ortográfico/fonológico pensamos que las tareas de clasificación y de pronunciación pueden funcionar suficientemente bien, más si cabe cuando las técnicas análogas que han empleado métodos de seguimiento ocular (v.g., técnica de *fast priming*; Sereno y Rayner, 1992) han dado lugar a resultados poco claros.

Por lo que respecta a la tarea "reina" en la investigación del acceso al léxico, la decisión léxica, pensamos que, una de sus variantes, la decisión léxica simple (*go/no-go lexical decision task*) debería ir reemplazando a la tarea de decisión léxica estándar, salvo que el investigador esté interesado en conocer los tiempos de reacción para las pseudopalabras. Como hemos indicado anteriormente, con la tarea de decisión léxica "simple" los tiempos de reacción son menores y la tasa de error también disminuye, aspectos ambos deseables en cualquier tarea experimental. Además, la tarea de decisión léxica simple puede aplicarse a poblaciones particulares, como los escolares, sin demasiados problemas en términos de porcentajes de errores. En cambio, la tarea de decisión léxica estándar con tales poblaciones da lugar a porcentajes de errores mucho mayores (Yelland, comunicación personal).

Finalmente, en la segunda sección nos hemos centrado en las estrategias de análisis de los datos experimentales y el modo más adecuado de seleccionar el material estimular. Es importante indicar que las normas de la APA (*American Psychological Association*) van siendo modificadas, dando prioridad a conceptos como el tamaño del efecto o los intervalos de confianza más que a un proceso ciego de contraste de hipótesis. No obstante, resulta claro que aún queda cierto camino por recorrer para los psicólogos experimentales a la hora de seleccionar y analizar los datos experimentales. Para concluir, aunque la predicción de Cutler (1981) respecto a que no podríamos efectuar experimentos con material lingüístico en los años 90 no se ha cumplido, sí es cierto que con el descubrimiento de nuevas variables que modulan el acceso al léxico va siendo cada vez más difícil diseñar experimentos que se mantengan a salvo de posibles variables extrañas.

ABSTRACT

Psychology of reading and visual word recognition: A review of experimental techniques and procedures for data analysis.

This paper offers a critical review of the methods used in psycholinguistic research, as well as an analysis of the different procedures for conducting data analyses and for selecting the items in the experiments. The major emphasis of the first part of the review is on the techniques used for studying visual word recognition (e.g., lexical decision, naming, etc.). Theoretical and practical considerations concerning the use of the different techniques are discussed. The second part of the review focuses on three aspects involved in process of planning and analyzing the experiments: (a) the logic of hypothesis testing, (b) the analysis of reaction time and the use of trimming procedures, and (c) the procedures for selecting items in word recognition experiments.

Key words: lexical- access, response-time, data-analysis.

REFERENCIAS

- Abelson. R. P. (1997). On the surprising longevity of flogged horses: Why there is a case for the significance test. *Psychological Science*, 8, 12-15.
- Alameda, J. R. y Cuetos, F. (1995). *Diccionario de frecuencia de las unidades lingüísticas del castellano*. Oviedo: Servicio de publicaciones de la Universidad de Oviedo.
- Algarabel, S., Meseguer, E., Dasí, C. y Ruiz, J. C. (1989). Los resultados nulos en la investigación experimental: Su incidencia y aplicaciones. *Psicológica*, 141 141-152.
- Andrews, S. (1989). Frequency and neighborhood effects on lexical access: Activation or search? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 15, 802-814.
- Andrews, S. (1997). The effects of orthographic similarity on lexical retrieval: Resolving neighborhood conflicts. *Psychological Bulletin and Review*, 4, 439-461.
- Anscombe, F. J. (1960). Rejection of outliers. *Technometrics*, 2, 123-147.
- Balota, D. A. y Chumbley, J. I. (1984). Are lexical decisions a good measure of lexical access? The role of word frequency in the neglected decision stage. *Journal of Experimental Psychology: Human Perception and Performance*, 10, 340-357.
- Balota, D. A. y Chumbley, J. I. (1990). Where are the effects of frequency in visual word recognition tasks? Right where we said they were! Comment on Monsell, Doyle, and Haggard (1989). *Journal of Experimental Psychology: General*, 119, 231-237.
- Baluch, B. y Besner, D. (1991). Visual word recognition: Evidence for strategic control of lexical and nonlexical routines in oral reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 17, 644-652.
- Barnett, V. y Lewis, T. (1995). *Outliers in statistical data*. New York: Wiley.
- Berkson, J. (1938). Some difficulties of interpretation encountered in the application of the chi-square test. *Journal of the American Statistical Association*, 33, 526-542.
- Berkson, J. (1942). Tests of significance considered as evidence. *Journal of the American Statistical Association*, 37, 325-335.

- Broadbent, D. E. (1967). Word frequency effect and response bias. *Psychological Review*, 74, 1-15.
- Carpenter, P. A. (1984). The influence of methodologies on psycholinguistic research: A regression to the Whorfian hypothesis. En D. Kieras y M. Just (Comps.), *New methods in reading comprehension research* (pp. 3-11). Hillsdale, NJ: Erlbaum.
- Carreiras, M. (1997). *Descubriendo y procesando el lenguaje*. Madrid: Trotta.
- Carreiras, M., García-Albea, J. E. y Sebastián, N. (1996). *Language processing in Spanish*. NJ: Erlbaum.
- Carreiras, M., Álvarez, C. J., & de Vega, M. (1993). Syllable frequency and visual word recognition in Spanish. *Journal of Memory and Language*, 32, 766-780.
- Catlin, J. (1969). On the word-frequency effect. *Psychological Review*, 76, 504-506.
- Clark, H. H. (1973). The language-as-a-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior*, 12, 335-359.
- Clifton, C., Speer, S. y Abney, S. P. (1991). Parsing arguments: Phrase structure and argument structure as determinants of initial parsing decisions. *Journal of Memory and Language*, 30, 251-272.
- Cohen, J. (1990). Things I have learned (so far). *American Psychologist*, 45, 1304-1312.
- Cutler, A. (1981). Making up materials is a confounded nuisance: or Will we be able to run any psycholinguistic experiments at all in 1990? *Cognition*, 10, 65-70.
- Danks, J. H. (1986). Identifying component processes in text comprehension: Comment on Haberlandt and Graesser. *Journal of Experimental Psychology: General*, 115, 193-197.
- de Partz, M. (1986). Re-education of a deep dyslexic patient: Rationale of the method and results. *Cognitive Neuropsychology*, 3, 159-180.
- de Vega, M. y Cuetos, F. (1999). *Psicolingüística del español*. Madrid: Trotta.
- Demonet, J. F., Price, C., Wise, R. y Frackowiak, R. S. J. (1994). A PET study of cognitive strategies in normal subjects during language tasks: Influence of phonetic ambiguity and sequence processing on phoneme monitoring. *Brain*, 117, 671-682.
- Falk, R. y Greenbaum, C.W. (1995). Significance tests die hard: the amazing persistence of a probabilistic misconception. *Theory and Psychology*, 2, 75-98.
- Fischler, I. (1977). Semantic facilitation without association in a lexical decision task. *Memory and Cognition*, 5, 335-339.
- Forster, K. I. (1979). Levels of processing and the structure of language processor. En W. E. Cooper y E. C. T. Walker (Comps.), *Sentence processing: Psycholinguistic studies presented to Merrill Garrett*. Cambridge, MA.: MIT Press.
- Forster, K. I. (1981). Priming and the effects of sentence and lexical contexts on naming time: Evidence for autonomous lexical processing. *Quarterly Journal of Experimental Psychology*, 33 A, 465-495.
- Forster, K. I. (1998). The pros and cons of masked priming. *Journal of Psycholinguistic Research*, 27, 203-233.
- Forster, K. I. (1999). *How should items be selected in word recognition experiments?* Enviado para publicación.
- Forster, K. I. y Shen, D. (1996). No enemies in the neighborhood: absence of inhibitory neighborhood effects in lexical decision and semantic categorization. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 22, 696-713.
- Forster, K. I. y Taft, M. (1994). Bodies, antibodies, and neighborhood density effects in masked form priming. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 20, 844-863.

- Frick, R. W. (1996). The appropriate use of null hypothesis testing. *Psychological Methods, 1*, 379-390.
- Gollan, T. H. y Frost, R. (1999). *The syntactic route to grammatical gender: Evidence from Hebrew*. Enviado para publicación.
- Gordon, B. (1983). Lexical access and lexical decision: Mechanisms of frequency sensitivity. *Journal of Verbal Learning and Verbal Behavior, 22*, 24-44.
- Gordon, B. y Caramazza, A. (1992). Lexical decision for open- and closed-class words: Failure to replicate differential frequency sensitivity. *Brain and Language, 15*, 143-160.
- Grainger, J. y Jacobs, A. M. (1996). Orthographic processing in visual word recognition: A multiple read-out model. *Psychological Review, 103*, 518-565.
- Grainger, J. y Seguí, J. (1990). Neighborhood frequency effects in visual word recognition: A comparison of lexical decision and masked identification latencies. *Perception and Psychophysics, 47*, 191-198.
- Greenwald, A. G., González, R., Harris, R. J. y Guthrie, D. (1996). Effects size and *p* values: What should be reported and what should be replicated? *Psychophysiology, 33*, 175-183.
- Haberlandt, K. (1994). Methods in reading research. En M. A. Gernsbacher (Comp.), *Handbook of psycholinguistics* (pp.1-31). San Diego, CA: Academic Press.
- Haberlandt, K. y Bingham, G. (1978). Verbs contribute to the coherence of brief narratives: Reading related and unrelated sentence triples. *Journal of Verbal Learning and Verbal Behavior, 17*, 419-425.
- Haberlandt, K. y Graesser, A. C. (1990). Integration and buffering of new information. En A. Graesser y G. Bower (Comps.), *Interferences and text comprehension* (pp. 71-87). San Diego, CA: Academic Press.
- Halle, M. (1990). Phonology. En D. N. Osherson y H. Lasnik (Comps.), *Language: An invitation to cognitive science* (pp. 43-68). Cambridge, MA: MIT Press.
- Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J. y Stahel, W.A. (1986). *Robust statistics: The approach based on influence functions*. New York: Wiley.
- Hang, T. H. (1996). Connectionist modeling of the recovery of language functions following brain damage. *Brain and Language, 52*, 7-24.
- Harlow, L. L., Mulaik, S. A. y Steiger, J. H. (1997). *What if there were no significance tests?* NJ: Erlbaum.
- Harris, R. J. (1991). Significance tests are not enough: The role of effect-size estimation in theory corroboration. *Theory and Psychology, 1*, 375-382.
- Heathcote, A. (1996). RTSYS: A DOS application for the analysis of reaction time data. *Behavior Research Methods, Instruments, and Computers, 28*, 427-445.
- Hino, Y. y Lupker, S. J. (1996). Effects of polysemy in lexical decision and naming: An alternative to lexical access accounts. *Journal of Experimental Psychology: Human Perception and Performance, 22*, 1331-1356.
- Hino, Y. y Lupker, S. J. (1998). The effects of word frequency for Japanese Kana and Kanji words in naming and lexical decision: Can the dual-route model save the lexical-selection account? *Journal of Experimental Psychology: Human Perception and Performance, 24*, 1431-1453.
- Huber, P. (1981). *Robust statistics*. New York: Wiley.
- Hudson, P. T. W. y Bergman, M. W. (1985). Lexical knowledge in word recognition: Word length and word frequency in naming and lexical decision tasks. *Journal of Memory and Language, 24*, 46-58.

- Just, M. A. y Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review*, 87, 329-354.
- Just, M. A., Carpenter, P. A. y Woolley, J. D. (1982). Paradigms and processes in reading comprehension. *Journal of Experimental Psychology: General*, 111, 228-238.
- Keppel, G. (1982). *Design and analysis. A researcher's handbook*. Prentice Hall.
- Kutas, M. y Hillyard, S. A. (1980). Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science*, 207, 203-205.
- Landauer, T. K. y Freedman J. L. (1968). Information retrieval from long-term memory: Category size and recognition time. *Journal of Verbal Learning and Verbal Behavior*, 7, 291-295.
- Luce, R. D. (1986). *Response times: Their role in inferring elementary mental organization*. New York: Oxford University Press.
- MacDonald, R. R. (1997). On statistical testing in psychology. *British Journal of Psychology*, 88, 333-347.
- Maxwell, S. E. y Delaney, H. D. (1990). *Designing experiments and analysis data. A model comparison perspective*. Belmont, CA: Wadsworth Publishing Company.
- McConkie, G. W., Hogaboam, T. W., wolverton, G. S., Zola, D. y Lucas, P.A. (1979). Toward the use of eye movements in the study of language processing. *Discourse processes*, 2, 157-177.
- Monsell, S. (1990). Frequency effects in lexical tasks: Reply to Balota and Chumbley. *Journal of Experimental Psychology: General*, 119, 335-339.
- Monsell, S. (1991). The nature and locus of word frequency effects in reading. En D. Besner y G. W. Humphreys (Comps.), *Basic processes in reading: Visual word recognition*. Hillsdale, NJ: Erlbaum.
- Monsell, S., Doyle, M. C. y Haggard, P. N. (1989). Effects of frequency on visual word recognition tasks: Where are they? *Journal of Experimental Psychology: General*, 118, 43-71.
- Oakes, M. (1986). *Statistical inference: A commentary for the social and behavioral sciences*. New York: Wiley.
- Patterson, K. E. (1981). Neuropsychological approaches to the study of reading. *British Journal of Psychology*, 72, 151-174.
- Perea, M. y Algarabel, S. (1999). *Puntuaciones atípica y potencia estadística con diferentes procedimientos de análisis de los tiempos de reacción: Un estudio de simulación*. Enviado para publicación.
- Perea, M. y Gotor, A. (1997). Associative and semantic priming effects occur at very short SOAs in lexical decision and naming. *Cognition*, 67, 223-240.
- Perea, M. y Pollatsek, A. (1998). The effects of neighborhood frequency in reading and lexical decision. *Journal of Experimental Psychology: Human Perception and Performance*, 24, 767-779.
- Perea, M., & Carreiras, M. (1998). Effects of syllable frequency and neighborhood syllable frequency in visual word recognition. *Journal of Experimental Psychology: Human Perception and Performance*, 24, 1-11.
- Perea, M., Rosa, E. y Gómez, C. (1999). *The word-frequency effect in the yes-no and in the go/no-go lexical decision task*. Enviado para publicación.
- Perfetti, C. A., Bell, L. y Delaney, S. (1988). Automatic phonetic activation in silent word reading: Evidence from backward masking. *Journal of Memory and Language*, 27, 59-70.

- Petersen, S. E., Fox, P. T., Posner, M. I., Mintum, M. y Raichle, M. E. (1989). Positron emission tomographic studies of the processing of single words. *Journal of Cognitive Neuroscience*, *1*, 153-170.
- Plaut, D. C. (1996). Relearning after damage in connectionist networks: toward a theory of rehabilitation. *Brain and Language*, *52*, 25-82.
- Pollatsek, A. y Well, A. (1995). On the use of counterbalanced designs in cognitive research: A suggestion for a better and more powerful analysis. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*, 785-794.
- Pollatsek, A., Perea, M. y Binder, K. (1999). The effects of neighborhood size in reading and lexical decision. *Journal of Experimental Psychology: Human Perception and Performance*, *25*.
- Pugh, K. R., Shaywitz, B. A., Shaywitz, S. E., Shankweiler, D. P., Katz, L., Fletcher, J. M., Skudlarski, P., Fulbright, R. K., Constable, R. T., Bronen, R. A., Lacadie, C. y Gore, J. C. (1997). Predicting reading performance from neuroimaging profiles: The cerebral basis of phonological effects in printed word identification. *Journal of Experimental Psychology: Human Perception and Performance*, *23*, 299-318.
- Ratcliff, R. (1979). Group reaction time distributions and an analysis of distribution statistics. *Psychological Bulletin*, *86*, 446-461.
- Ratcliff, R. (1993). Methods for dealing with reaction time outliers. *Psychological Bulletin*, *114*, 510-532.
- Rayner, K. (1978). Eye movements in reading and information processing. *Psychological Bulletin*, *85*, 618-660.
- Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, *124*, 372-422.
- Rayner, K. y Pollatsek, A. (1989). *The psychology of reading*. Englewood Cliffs, NJ: Prentice-Hall.
- Rayner, K., Sereno, S. C., Morris, R. K., Schmauder, A. R. y Clifton, C. J. (1989). Eye movements and on-line language comprehension processes. *Language and Cognitive Processes*, *4*, SI21-SI49.
- Reichle, E. D., Pollatsek, A., Fisher, D. L. y Rayner, K. (1998). Toward a model of eye movement control in reading. *Psychological Review*, *105*, 125-157.
- Rubinstein, H., Garfield, L. y Millikan, J. A. (1970). Homographic entries in the internal lexicon. *Journal of Verbal Learning and Verbal Behavior*, *9*, 487-494.
- Rubinstein, H., Lewis, S. S. y Rubinstein, M. A. (1971a). Homographic entries in the internal lexicon: Effects of systematicity and relative frequency of meanings. *Journal of Verbal Learning and Verbal Behavior*, *10*, 57-62.
- Rubinstein, H., Lewis, S. S. y Rubinstein, M. A. (1971b). Evidence for phonemic recoding in visual word recognition. *Journal of Verbal Learning and Verbal Behavior*, *10*, 645-657.
- Rugg, M. D. y Nagy, M. E. (1987). Lexical contributions to nonword-repetition effects: Evidence from event-related potentials. *Memory and Cognition*, *15*, 473-481.
- Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods*, *1*, 115-129.
- Sebastián-Gallés, N. (1991). Reading by analogy in a shallow orthography. *Journal of Experimental Psychology: Human Perception and Performance*, *17*, 471-477.
- Sereno, S. C. y Rayner, K. (1992). Fast priming during eye fixations in reading. *Journal of Experimental Psychology: Human Perception and Performance*, *18*, 173-184.

- Singer, M. (1990). *Psychology of language: An introduction to sentence and discourse processes*. Hillsdale, NJ: Erlbaum.
- Snodgrass, J. G. y Mintzer, M. (1993). Neighborhood effects in visual word recognition. *Memory and Cognition*, 21, 247-266.
- Tabossi, P. y Laghi, L. (1992). Semantic priming in the pronunciation of words in two writing systems: Italian and English. *Memory and Cognition*, 20, 303-313.
- Taft, M. (1991). *Reading and the mental lexicon*. London: Erlbaum.
- Ulrich, R. y Miller, J. (1994). Effects of outlier exclusion on reaction time analysis. *Journal of Experimental Psychology: General*, 123, 34-80.
- Van Petten, C. y Kutas, M. (1987). Ambiguous words in context: An event-related potential analysis of the time course of meaning activation. *Journal of Memory and Language*, 26, 188-208.
- Wilcox, R. R. (1998). How many discoveries have been lost by ignoring modern statistical methods? *American Psychologist*, 53, 300-314.

(Revisión aceptada: 24/5/99)