

ECUACIONES DE PREDICCIÓN DE LECTURABILIDAD

JOSÉ LUIS RODRÍGUEZ DIÉGUEZ
PILAR MORO BERIHUETE
MARIVÍ CABERO PÉREZ
*Dptº de Didáctica, Organización
y Métodos de Investigación.
Universidad de Salamanca*

RESUMEN

Los autores dan cuenta de una investigación efectuada en la provincia de Salamanca con una muestra total de 200 textos, que fueron valorados en su comprensión por un total de 1.600 alumnos de diversos niveles de enseñanza, desde cuarto curso de E.G.B. hasta cuarto curso de estudios universitarios.

La comprensión fue inicialmente evaluada por medio de un total de 10 variables, que luego se factorizaron, obteniéndose dos puntuaciones factoriales por cada texto como indicadores de comprensión.

A partir de esta muestra, y considerando un conjunto de predictores lingüísticos de índole superficial, se determinan las ecuaciones de predicción para cada nivel de enseñanza y en los dos factores detectados. La correlación múltiple obtenida mediante las distintas ecuaciones de predicción varía, para el primer factor, entre .590 y .760, y para el segundo factor entre .197 y .577.

PREDICTION OF READABILITY EQUATIONS

SUMMARY

The authors report on research carried out in the province of Salamanca with a total sample of 200 texts, which were evaluated as to their comprehension by a total of 1,600 students from different levels of education, ranging from fourth year of E.G.B.* to fourth year of University studies.

* E. G. B. comprises the first eight years of formal schooling in the current Spanish system (ages 6-13).

Comprehension was first evaluated by means of a total of 10 variables, which were then factorized, thus obtaining two factorial scores for each text as comprehension indicators.

From our sample, and taking into consideration a set of linguistic predictors of a superficial nature, the prediction equations were determined for each level of education and in the two factors detected. The multiple correlation obtained by the different prediction equations varied, for the first factor, between .590 and .760, and for the second factor, between .197 and .577.

EQUATIONS DE PREDICTION DE LECTURABILITE

RESUME

Les auteurs présentent une recherche menée dans la province de Salamanca sur un échantillon de 200 textes, évalués sur leur compréhension par un total de 1.600 élèves de divers niveaux d'enseignement, depuis la 4^e année de l'E.G.B. (*école primaire*) jusqu'à la 4^e année des études universitaires.

La compréhension a été initialement évaluée au moyen de 10 variables par la suite factorisées, obtenant deux valeurs factoriels par texte comme indicateurs de compréhension.

A partir de cet échantillon, et considérant un ensemble de prédicteurs linguistiques de type superficiel, on a déterminé les équations de prédiction pour chaque niveau d'enseignement et dans les deux facteurs observés. La corrélation multiple obtenue par les différentes équations de prédiction varie, pour le premier facteur, entre .590 et .570 et, pour le deuxième facteur, entre .197 et .577.

1. UNA PRECISIÓN INICIAL

Las "jergas" científicas suelen responder a necesidades de matización entre territorios o conceptos muy próximos. A este criterio responde la distinción que se realiza en inglés entre "legibility" y "readability". Los equivalentes castellanos que cabría considerar serían "legibilidad" y "lecturabilidad".

La legibilidad es una característica morfológica de un texto, alude a aspectos relativos a su forma externa: tipos, tamaños, interlineado, etc. Los estudios de Tinker (1963) son, tal vez, los más interesantes en torno a este aspecto. Es incuestionable que ciertos tipos son más claros y legibles que otros, y que la longitud total de la línea, el tamaño de la página o el de las letras puede influir facilitando la lectura y la comprensión de un texto.

La lecturabilidad se refiere a unas características de índole lingüística. Hace referencia sobre todo a aspectos léxicos, a construcciones gramaticales, a signos de puntuación. Y es evidente que estos rasgos inciden también sobre la posibilidad de comprensión.

Como consecuencia cabe afirmar que un texto manuscrito suele ser menos legible que uno impreso. Y un cuento de Gloria Fuertes tiende a ser más lecturable que un texto de Hegel.

Nuestro trabajo se centra en la valoración de la lecturabilidad de los textos, en la predicción de la comprensión de dichos textos a partir de indicadores lingüísticos superficiales.

2. LOS ESTUDIOS SOBRE LECTURABILIDAD EN CASTELLANO

Las fórmulas de lecturabilidad en cualquier idioma no pueden importarse sin más. Ni siquiera son aceptables plenamente aquellas adaptaciones “por analogía” como las que efectúan por ejemplo Kandel y Moles con relación al francés (1958). Los estudios sobre lecturabilidad pueden replicarse, pero nunca “traducirse”.

La historia de las fórmulas de lecturabilidad en castellano es curiosa. Por ejemplo: las más frecuentes y citadas suelen ser las aportaciones procedentes de autores no hispanohablantes.

Las primeras fórmulas conocidas son las de Spaulding (1951, 1956), que hace uso de dos variables para efectuar la predicción: la longitud media de la frase y la usualidad de vocabulario, apoyándose en el diccionario de Buchanan (1941).

Fernández Huerta, precursor en este tema como en otros muchos, comienza a trabajar sobre la lecturabilidad en torno a 1959.

Trabajos posteriores sobre el tema son los de Gutiérrez (1972), Patterson (1972), Thonis (1976), García (1977), Gillian, Peña y Mountain (1980), Vari-Cartier (1981) y Cradford (1984).

Por nuestra parte, este tema se planteó en los años setenta en el Departamento de Didáctica de la Universidad de Valencia, donde se efectuó el primer trabajo de López Rodríguez (1976), que da lugar posteriormente a los de 1981 y 1982, dirigidos todos ellos por Rodríguez Diéguez.

Esta línea de trabajo se traslada posteriormente a la Universidad de Salamanca, en cuyo Departamento de Didáctica se efectúan una serie posterior de trabajos en esta misma línea (Rodríguez Diéguez, 1983, 1988, 1989).

Una interesante y sintética perspectiva de todos estos estudios aparece en el documentado estudio de Annette T. Rabin (1988).

Estos trabajos, en la Universidad de Salamanca, se cierran por ahora con los de Moro (1991) y Cabero (1991).

3. LAS VARIABLES DE LOS ÚLTIMOS ESTUDIOS

Los dos últimos estudios mencionados se han efectuado con una metodología similar, con una muestra amplia de sujetos de cada uno de los niveles estudiados, y considerando una serie de indicadores de comprensión lectora diferentes.

La variable dependiente del estudio, que más adelante analizaremos, ha sido la facilidad de lectura de los textos estudiados, evaluada por medio de:

a) Diversas variantes de la prueba Cloze aplicados a muestras de sujetos de distintas edades y situaciones culturales.

b) Valoración de la dificultad estimada en cada uno de los textos, efectuada por un total de diez jueces, todos ellos profesores.

c) Pruebas convencionales de comprensión (redacción, enumeración de ideas principales, etc.).

Las variables independientes venían moduladas por la exigencia del trabajo. Habían de ser aspectos expresos en el texto, determinables de modo objetivo, y cuantificables. Se buscaba una ecuación de predicción. La única vía susceptible de arrojar resultados válidos había de responder a esta característica. Las posibilidades de otras vías de acceso a la evaluación de la comprensión no quedan, ni mucho menos, cerradas en este planteamiento (Rodríguez Diéguez y Beltrán de Tena, 1985; Rodríguez Diéguez, 1991).

Vamos a estudiar ahora las variables predictoras.

3.1. Descripción de las variables independientes

Se utilizaron las siguientes variables:

Variable 1. Pronombres.

Total de pronombres que aparecen en el texto, expresado cuantitativamente mediante la siguiente fórmula:

$$\frac{\text{n pronombres 1 y 2 persona}}{\text{n total de palabras del texto}} \times 1000$$

Variable 2. Vocabulario G-H.

Usualidad del vocabulario, utilizando como referente el vocabulario usual de García Hoz (1953), convertido en indicador a partir de la siguiente transformación:

$$\frac{\text{palabras total del texto} - \text{palabras del diccionario}}{\text{palabras total del texto}} \times 1000$$

Variable 3. Desviación típica.

El valor de la desviación típica de la distribución de letras por palabra de cada uno de los textos.

Variable 4. Vocabulario TV.

Usualidad del vocabulario del texto, evaluada con referencia al vocabulario usual obtenido por Lorenzo Delgado (1981) a partir de una muestra de programas de televisión, tratado a través de la misma fórmula que en la variable 2 fórmula

Variable 5. Palabras por frase.

Proporción de palabras por frase en el total del texto que se evalúa.

$$\frac{\text{n pal. totales del texto}}{\text{n. de frases}} \times 1000$$

Variable 6. Total de puntos.

Se expresaba mediante la fórmula:

$$\frac{\text{n puntos totales del texto}}{\text{n palabras total del texto}} \times 1000$$

Variable 7. Total puntos y aparte.

Se utilizó la fórmula:

$$\frac{\text{n total puntos y aparte}}{\text{n. total palabras texto}} \times 1000$$

Variable 8. Comas.

Se calculó el indicador a partir de:

$$\frac{\text{n comas totales del texto}}{\text{n palabras total texto}} \times 1000$$

Variable 9. Media más 2.58 sigmas.

Se basa en la determinación de la media de letras por palabras en cada uno de los textos, a la que se le suma 2.58 veces el valor de la desviación típica de esa misma distribución.

Variable 10. Puntos y coma.

Este índice se calcula por medio de:

$$\frac{\text{n puntos y coma totales texto}}{\text{n total palabras texto}} \times 1000$$

Variable 11. Nombres propios.

Estimado a través del número de mayúsculas no justificadas ortográficamente que aparecen en el texto, y tratado con la fórmula:

$$\frac{\text{n mayúsculas- n total puntos}}{\text{n total palabras texto}} \times 1000$$

Variable 12. Tasa de redundancia modificada.

El cálculo habitual del indicador TTR suele efectuarse directamente sobre el texto original. En este caso hemos considerado las formas lexemáticas, sin morfemas. La fórmula utilizada ha sido:

$$\frac{\text{n pal. distintas del texto}}{\text{n total palabras texto}}$$

3.2. La descripción factorial de las variables independientes

Los resultados de estas doce variables iniciales, valoradas en todos y cada uno de los 200 textos con los que se trabajó, se sometieron a un análisis factorial por medio del procedimiento de Componentes Principales, y se procedió después a rotación con la técnica Varimax.

La matriz factorial rotada es la siguiente:

	Factor 1	Factor 2	Factor 3
1. Pronom. Pers.	,435	,255	,136
2. Voc. García Hoz.	,857	,012	-,041
3. σ palabras	-,598	-,13	-,386
4. Voc. TV L. Del.	,862	,009	-,003
5. Pal. por frase	,02	-,566	-,519
6. Total puntos	,344	,713	,062
7. Puntos aparte	,125	,788	-,042
8. Total comas	,02	,009	,645
9. $X + 2.58 \sigma$	-,734	-,202	-,337
10. Puntos coma	,125	-,217	,674
11. Nom. propios	,02	,573	-,169
12. TTR	-,75	-,001	,068

Se pretende con este análisis poner de manifiesto las características de las distintas variables consideradas como predictoras.

El agrupamiento en factores es coincidente con otros resultados que hemos obtenidos en otros momentos (Beltrán de Tena, 1983, Rodríguez Diéguez, 1991). Se presenta como una estructura simple claramente invariante. Los factores que la componen son los siguientes:

3.2.1. El factor lexicográfico

Se detecta la presencia de un factor inicial, de índole claramente lexicográfico, y que gira en torno a una serie de características de las palabras. Tasa de redundancia, número de pronombres, usualidad de vocabulario —tanto a partir del vocabulario de García Hoz como del de Lorenzo Delgado— y unas curiosas variables: desviación típica de la distribución de letras por palabra y media más 2,58 sigmas. Será necesario extendernos en torno a este factor.

Este factor I se estructura, como ya se ha dicho, alrededor del concepto de palabra. El concepto de palabra es notablemente ambiguo, como puede observarse en Ullman (1970) o en Martinet (1974).

Desde nuestra perspectiva, y a efectos puramente funcionales, hemos considerado la palabra como “un grupo de signos separado de los grupos vecinos por un blanco o una puntuación” (Muller, 1973).

La descripción de las variables implicadas en este primer factor parece subrayar un carácter altamente superficial. Usualidad, redundancia, longitud total del texto, longitud de palabras son atributos que, inicialmente, parecen no tener otro referente que características absolutamente formales, casi “morfológicas”.

Las variables “desviación típica” y “media más 2,58 sigmas” son las que, inicialmente, resultan más disonante con el resto de las propuestas. Un análisis inicial del sentido pleno de la segunda apunta hacia la longitud de las palabras utilizadas en el texto. La distribución de letras por palabras de un fragmento permite valorar la media de letras de ese texto. La utilización de la medida de dispersión nos permite estimar el intervalo de longitud máxima probable en un uno por cien de casos. Media de letras por palabra más/menos 2,58 sigmas marca el intervalo confidencial del 99 %. Si se considera sólo una cola de la distribución, el nivel de confianza se incrementa al 99,5 %. De este modo, tan solo cabe esperar que una palabra de cada doscientas tenga esa longitud. Se trata de una extrapolación de la longitud máxima esperable con esa probabilidad en ese texto concreto. Se podría decir que este indicador es una estimación de la máxima longitud esperable en las palabras utilizadas en un texto. Si un texto tiene como promedio de letras por palabra un valor de 8, y como desviación típica .94, el valor de 10.42 indicará la longitud máxima esperable tan sólo en el 0,5 % de palabras.

La longitud de palabra ha sido un indicador frecuentemente utilizado en las fórmulas de lecturabilidad en distintos idiomas, utilizada a partir de distintas valoraciones. La valoración de la longitud de las palabras es uno de los indicadores más clásicos en el ámbito de la lingüística cuantitativa. Y el indicador de longitud de palabras “media más 2,58 sigmas”, además de ser coherente, demuestra ser una estimación eficaz y más precisa que todas las efectuadas hasta ahora.

En las revisiones bibliográficas que hemos efectuado no hay referencia alguna a este indicador. Parece que su utilización no se ha producido en ninguna lengua. Es una aportación original de los estudios de Rodríguez Diéguez (1983, 1988, 1989, 1991, entre otros). Un estudio más completo de algunos aspectos de esta variable desde la perspectiva de las condiciones de producción puede verse en Herrera García (1992).

A partir de la Ley de Zipf (1949) podemos entrar en la otra perspectiva que aparece enlazada en nuestro factor I, relacionada con su usualidad. Entre ambas, ya se ha dicho, existe una estrecha concomitancia.

La primera constatación aparece por la presencia de una y otras variables en el mismo factor. Pero aún más expresivo es el caso de los estudios efectuados a partir de este estimador en otras lenguas. En portugués, en concreto, el uso del vocabulario del INIC (1984) demostró ser poco eficaz como predictor en análisis lingüísticos que ya habíamos trabajado en castellano. Los criterios utilizados para determinar la usualidad en este diccionario no eran homogeneizables con los diccionarios españoles anteriormente citados. En portugués, y según nuestros datos, este diccionario no “funciona”. La ausencia de capacidad predictiva del vocabulario usual en los análisis produjo un notable incremento de la variable “media + 2,58 sigmas” (Rodríguez Diéguez y Lebre Aires, 1990). Alertados sobre este resultado, se ha constatado igualmente que la supresión de las variables de usualidad en el proceso de obtención de la ecuación de predicción de lecturabilidad por el procedimiento “paso a paso” supone un incremento notable del poder predictivo de la variable de longitud a la que nos estamos refiriendo.

Pero todavía hay un tercer aspecto que resaltar de este indicador. Ya decimos que proporciona una buena valoración de la longitud de palabras, y que se asocia íntimamente con la usualidad de vocabulario. El tercer aspecto hace referencia a esta asociación con el vocabulario usual.

Frecuentemente se contraponen la usualidad, como resultado cuantitativo y superficial, a otra serie de constructos tal vez más elaborados. Así lo hace, por ejemplo, Hewitt (1980). La información poseída por el lector, sus esquemas y conceptos previos, son los auténticos determinantes de la comprensión según este autor.

Desde mi punto de vista, el fenómeno detectado puede tener otras lecturas. El vocabulario usual del escolar salmantino, o el del adulto de Villanueva de la Serena no es otra cosa que el inventario de términos más frecuentemente utilizado por el colectivo de niños residentes en Salamanca, o de adultos de aquella localidad extremeña. Y esos términos responden a la presencia de una serie de experiencias comunes a unos o a otros. Lo que diferenciará uno y otro vocabulario será precisamente aquellos objetos, aquellas realidades o aquellos acontecimientos que son frecuentes en uno y otro entorno. La palabra "hornazo" es posible que apareciera en el nivel usual en uno y otro contexto: alude a la torta hecha con embutido, huevos y masa de pan, que suele consumirse después de la cuaresma. Es una realidad, una experiencia previa que cabe pensar que sea común al niño salmantino, que sale el Lunes de Aguas a tomar el hornazo al campo, y al adulto de Villanueva de la Serena, que lo consume el Lunes de Pascua, el "día de la jira". Pero la palabra "candar" que el Diccionario de la Real Academia define como "cerrar con llave; por extensión, cerrar de cualquier modo" se entiende, en esta segunda acepción, de modo directo por cualquier chaval salmantino. "Se candó el río", se dice cuando las bajas temperaturas hacen que el hielo cierre la superficie del Tormes. No sería esperable que este sentido de "candar" no es usual. Y un adulto extremeño de la vegas del Guadiana entiende que "manear" o "amanear" consiste en poner cuerdas para atar las manos de un animal. Este término no es usual, sin embargo para adultos ni para niños salmantinos.

Pretendo con ello subrayar que el vocabulario usual de un colectivo no es sino un inventario de los tópicos, de las informaciones y de las experiencias poseídas y así denominadas por ese colectivo al que se aplica. Anderson y Freebody (1979) han mantenido la hipótesis del conocimiento del vocabulario como un indicador de la importancia de los esquemas y las estructuras conceptuales previas. Y en la misma línea se manifiesta Spiro (1980).

De algún modo, la hipotética "superficialidad" de la variable y del factor queda notablemente superada. La longitud de palabras y la usualidad del vocabulario se convierten así en variables con más carga cognitiva que la que inicialmente se suponía.

3.2.2. El factor fraseológico

El segundo factor definido en la estructura simple procedente del análisis factorial de las variables independientes podemos denominarlo factor fraseológico. Aparece definido por la conjunción de tres variables cuya concomitancia parece evidente: puntos y aparte, total de puntos, palabras por frase. Resulta menos explicable la presencia de la variable "nombres propios".

La variable "palabras por frase" correlaciona en sentido contrario a las otras dos.

El análisis de este factor, desde nuestra perspectiva, debe hacerse en conjunción con el factor III, conceptualmente muy próximo a él. Esa proximidad viene marcada por la aparición de saturaciones exclusivamente asociadas a las variables relacionadas con los signos de puntuación en uno y otro caso.

Sin embargo, esa proximidad conceptual no se corresponde con una relación estadística. Inicialmente, la rotación es ortogonal, por tanto no existe intercorrelación entre los factores. Pero además, en un segundo análisis del conjunto de datos, se efectuó una rotación oblicua. Los factores se mantienen nítidamente. Y la correlación entre los factores II y III es tan sólo de .162.

3.2.3. El factor de dificultad sintáctica

Este tercer factor está constituido por las variables “palabras por frase”, “puntos y coma” y “comas”. La correlación entre la primera y las dos últimas aparece con sentido contrario.

Uno y otro factor responden a la estructuración del texto en frases. Y el concepto de frase es suficientemente impreciso y arduo.

El problema queda en parte asociado a una dimensión de puntuación. Porque frases, párrafos, oraciones o sintagmas tienen un correlato relativamente estable con los signos de puntuación. Plano que no aparece suficientemente claro tampoco en esta otra perspectiva. La puntuación supone la presencia de una serie de signos. Y los signos de puntuación tienen una función específica: “encuadrar construcciones independientes, oraciones o estructuras insertas en la oración, proposiciones” (Salvador Mata, 1986).

El sentido relativamente nítido del punto y aparte se va progresivamente desdibujando al considerar el punto y seguido o el punto y coma. Se entra en una zona de matices cuyo correlato, frecuentemente estilístico, impide efectuar precisiones mayores. Pero desde nuestra perspectiva es conveniente señalar que la presencia de la variable “punto” y “punto y aparte” en un factor, frente a “coma” y “punto y coma” en el otro tiene que ver con esa nitidez que se alcanza a captar en la diferenciación progresiva entre los distintos signos de puntuación.

De aquí que la ruptura del texto en una serie de unidades de contenido suficientemente diferenciadas —caso del punto y seguido o el punto y aparte— se proyecte de un modo diferenciado con relación a la estructuración matizadamente presentada a través de los modos más delicados de presentar la información, tales como son las comas y los puntos y coma. La línea divisoria entre la diferenciación “fuerte” y la matizada puede encontrarse en la frontera entre el punto por una parte y el punto y coma por la otra.

4. LA PREDICCIÓN DE LA LECTURABILIDAD

Las fórmulas de lecturabilidad que se van a presentar ahora se basan en estos presupuestos. De su necesidad en castellano da buena razón la existencia de fórmulas equivalentes en casi todas las lenguas. Y de su futuro habla de modo suficientemente elocuente la interesante línea de estudios sobre legibilidad y lecturabilidad transferida a la pantalla del computador (Grupo de investigaciones CFPJ, 1987),

como consecuencia de la necesidad manifestada por la utilización, cada día más frecuente, del terminal pantalla como modo de presentación de información escrita.

La intención de nuestro estudio es efectuar la predicción de la comprensibilidad general de un texto a partir de los indicadores anteriormente señalados.

Para ello hemos de proceder a la determinación de una ecuación de predicción entre los valores de las mencionadas variables, que constituyen el conjunto de predictores, y unas pruebas de comprensión cumplimentada por una serie de sujetos como criterio.

4.1. *Valoración de la comprensión de los textos*

Los textos que constituyeron la muestra fue un conjunto de doscientos fragmentos literarios en castellano, y caracterizados por tres criterios básicos: una dificultad inicialmente estimada como aproximadamente similar, cierre temático y coherencia expresiva, y como final una extensión que oscilaba entre doscientas a quinientas palabras.

A partir de estos doscientos textos se prepararon las siguientes pruebas de comprensión:

1. *Una variante del test cloze*, estudiado inicialmente en el Departamento de Didáctica de la Universidad de Salamanca. Datos sobre la validez de esta prueba en castellano y con muestras españolas pueden verse en López Rodríguez (1981), Bayón Alvaro (1984), Quintero Gallego (1985), Monsalvo Díez (1985) y Rodríguez Diéguez (1991).

La muestra de nuestro estudio se seleccionó a partir de la población de la provincia de Salamanca.

Se dividió en tres grupos (rural, semi-rural y urbano) partiendo de los datos proporcionados por el Departamento de Geografía Física de la Universidad de Salamanca. Cada texto debía ser cumplimentado, en aquellas muestras que lo permitieran, por cinco sujetos de población urbana, cuatro de población rural y uno de población semi-rural, en proporción a la tasa de habitantes de cada zona.

A su vez se seleccionaron los cursos que compondrían la muestra. En Educación General Básica se eligieron los cursos de segundo, cuarto, sexto y octavo por dos razones fundamentales:

- a) No se podía aplicar la prueba a sujetos de primer curso dado que en muchos centros es en este momento cuando se inicia el aprendizaje de la lecto-escritura.
- b) Los avances conseguidos en la lectura serían más evidentes si se alternaban los cursos que si se llevaba a cabo una aplicación de la prueba en cursos sucesivos. Por ello al tomarse segundo curso como punto de partida, los restantes cursos serían todos los pares, cuarto, sexto y octavo.

No obstante, en la corrección de la prueba de comprensión el segundo curso de Educación General Básica tuvo que ser eliminado dada la dificultad de la prueba para este nivel, quedando la muestra en Educación General Básica formada por los cursos cuarto, sexto y octavo.

En el caso de la Formación Profesional, dado que la población escolar en esta modalidad educativa tiene un número inferior de estudiantes, formaron parte de la muestra todos los cursos y ramas, dividiéndola en los dos grandes bloques en que

se encuentra estructurada organizativamente: Formación Profesional de primer grado y Formación Profesional de segundo grado. La población escolar en estos niveles tuvo que ser agrupada en otros dos grandes bloques: población urbana y semi-rural, debido a la inexistencia de centros de este tipo en zonas rurales.

La muestra de Bachillerato estaba constituida por alumnos de segundo curso y C.O.U.

En enseñanza universitaria se trabajó con alumnos de segundo y cuarto curso de Derecho y Filosofía y Ciencias de la Educación.

Seleccionada la muestra y preparados los textos se aplicó la prueba a un total de diez alumnos por texto y nivel en proporción a cada tipo de población, obteniéndose dos mil protocolos por curso, un total de dieciocho mil protocolos en la investigación.

La corrección de la prueba de comprensión, nos permitió la obtención de diez variables criterio, que tendrían que ponerse en relación con las doce variables independientes obtenidas del análisis de los textos en su versión íntegra.

En la corrección de la prueba cloze solamente se aceptaron como válidas aquellas palabras que igualasen exactamente a la suprimida, no admitiéndose los sinónimos.

Como excepción, se valoraron positivamente algunas variantes del texto en casos tales como la introducción de un tiempo verbal diferente, siempre que coincidiese con el verbo suprimido; la supresión o inclusión de aumentativos y diminutivos; el cambio de singular a plural o viceversa, y el caso en el que se completaba el espacio en blanco con dos palabras, cuando una de ellas corresponde a la original suprimida.

A partir de los mismos protocolos, se procedió a obtener las siguientes puntuaciones cloze:

1. Porcentaje total de aciertos sobre el número de espacios en blanco multiplicados por cien. Con esta primera forma de corrección se pretendía neutralizar la longitud del texto, ya que no todos tenían el mismo número de palabras.

2. Otra vía de homogeneización de los resultados consistió en "cortar" todos los textos, a efectos de evaluación, para que tuvieran la misma longitud. Para ello se valoraron un total de 49 items cloze, número de elementos que aparecían en el texto más corto.

3. La tercera puntuación asignada fue la puntuación directa del cloze, es decir el total de aciertos obtenidos por el sujeto en la prueba, independientemente de la longitud del texto.

4. La puntuación directa del cloze menos las dos primeras líneas; con ello se pretendía conseguir un cierto entrenamiento de los sujetos que podría influir en el número de aciertos conseguidos.

2. Ideas principales del texto. Era la segunda prueba de comprensión, en la cual se pedía a los sujetos, una vez finalizada la realización de la prueba cloze, que escribiesen las tres ideas que considerasen principales de la lectura.

A fin de evitar una valoración excesivamente subjetiva se intentó el acercamiento, siquiera fuera elemental, a un análisis de contenidos comparado entre texto original y la propuesta de los sujetos. Para ello se procedió a la extracción de las tres o cuatro ideas principales de cada texto ordenándolas según su importancia. A partir de estos datos se valoraban los protocolos.

Como consecuencia los doscientos textos, en todos y cada uno de los niveles, contaban con un total de diez puntuaciones conjuntos de puntuaciones. La valora-

ción que se atribuía en cada una de las escalas a cada texto era la puntuación media obtenida por los diez alumnos que habían respondido a la prueba cloze a la enumeración de ideas.

A fin de evitar el posible peso excesivo de puntuaciones extremas, se realizó otro cómputo más: en cada prueba se eliminaron la puntuación mayor y menor de la serie de diez. Dio lugar a una serie paralela de puntuaciones.

3. *Jueces*. Todavía se consideró un tercer tipo de prueba de comprensión. Consistió en la lectura de los doscientos textos por diez jueces elegidos al azar. Éstos tenían que asignar una puntuación a cada texto, que oscilaba entre cero y diez, en función de su dificultad o facilidad. Cuanto menor fuese la puntuación asignada a un texto mayor sería su dificultad y viceversa.

4.2. *Análisis estadístico*

Una vez determinadas las variables se procedió a realizar una serie de análisis factoriales. Con ello se pretendía conseguir:

- a) La obtención de la naturaleza de las construcciones factoriales de las variables estudiadas.
- b) Estudiar la posibilidad de combinar las puntuaciones criterio de un modo óptimo mediante la consideración de las puntuaciones factoriales.

De este modo los factores obtenidos nos van a describir, por un lado la estructura de los criterios utilizados y por otro nos van a precisar los criterios de predicción con las puntuaciones factoriales correspondientes.

Un primer análisis global nos arrojó los dos factores siguientes:

Factor I o factor *cloze*, ya que en él se agrupan todas las variables relativas al test cloze. Satura todas aquellas variables que implican derivar información del contexto general, así pues, cuanto más familiar le sea éste al sujeto mayores serán sus posibilidades de comprender el mensaje.

Factor II o factor *ideas-jueces* donde se agrupan las variables de las ideas principales del texto y las valoraciones de los jueces. Es un factor que combina un determinado conocimiento lingüístico y una valoración global de su dificultad, ya que supone la expresión escrita de lo que se considera más importante, siendo destacable sobre todo en los cursos inferiores donde la producción escrita de los sujetos es bastante limitada. El juicio promedio de los jueces se asocia a la variable ideas.

Este análisis global se vio confirmado esencialmente en su invarianza en los posteriores análisis factoriales por curso, en los que, obviamente, hubo que prescindir de la variable criterio "jueces".

A partir de los diversos análisis factoriales por cursos se obtuvieron las puntuaciones factoriales de cada texto-sujeto en cada uno de los dos factores obtenidos. Estas puntuaciones factoriales, que combinaban los diversos criterios estudiados, se utilizaron como puntuaciones criterio para la obtención de la ecuación de predicción.

4.3. *Las ecuaciones de predicción*

Los índices de correlación obtenidos denotan una relación sustancial, en el primer factor, mientras que en el segundo factor se obtiene una relación mucho más moderada.

Es curioso observar cómo la línea de intendidad en las correlaciones va descendiendo progresivamente. Los resultados de los alumnos universitarios son sensiblemente inferiores a los de los alumnos de Educación General Básica.

En el factor II, y en Formación Profesional 1, en 2º de BUP y en 4º de estudios universitarios la correlación no es significativa.

Posibles explicaciones de esta situación podrían ser, entre otras, las siguientes:

a/ La falta de interés de los alumnos en la cumplimentación de la prueba y/o,

b/ Que los criterios de evaluación de las ideas principales no fuesen todo lo objetivos que deberían haber sido, produciéndose una falta de concomitancia entre lo que el alumno y el evaluador consideran importante.

Los resultados podemos verlos a continuación en la tabla 2:

COEFICIENTES DE CORRELACION

CURSO/CRITERIO	Factor CLOZE	Factor IDEAS
4º	.734	.577
6º	.700	.417
8º	.773	.477
FP1	.760	---
FP2	.749	.418
2ºBUP	.726	---
COU	.682	.311
Univ.2º	.626	.197
Univ.4º	.590	---

Es obvio que el Factor Cloze ejerce un mayor peso que el Factor Ideas principales del texto, no obstante consideramos de interés el exponer ambas ecuaciones a continuación:

Ecuaciones de predicción del factor I de comprensión lectora.

4 ^e	$ID = 5.321 + .049X7 - .319X0 - .457X9 + .332X11$
6 ^e	$ID = 4.423 + 1.170X7 + .416X9$
0 ^e	$ID = 1.199 - .036X3 + .137X4 - .3662e-3 + .012X7$
FPI	$ID = 1.436 - 1.048X3 + .151X4 + .773X7 - .404X8 + .523X11$
FP2	$ID = .725 - .834X3 + 151 X4 + .844X7 - .251X8 - 1.137X10 + .581X11$
BUP	$ID = 1.9 - .691 X3 - 4.059X12 + .233X4 + .367X6 - .207X8 - 1.219X10 + .509X11$
COU	$ID = 3.101 - .801X3 - .005X5 - 3.657X12 + .203X4 - .494X8 + .571X11$
Un2	$ID = 3.012 - .715X3 - .004X5 - 3.662X12 + .177X4 - .382X8 - 1.628X10 + .524X11$
Un4	$ID = 1.073 - .005X5 - .3.352X4 - .549X8 + .715X11$

Ecuaciones de predicción del factor II de comprensión lectora

4º	$ID = 2.104 + .522X6 - 4.346X12$
6º	$ID = -.229 + .661X7$
8º	$ID = -1.539 + 1.02X2$
FP1	SIN ECUACION
FP2	$ID = 2.11 - 3.536X12$
BUP	SIN ECUACION
COU	$ID = -2.19 + .27X9 - .399X11$
Un2	$ID = 1.3 + .006X5$
Un4	SIN ECUACION

Una visión de conjunto de las ecuaciones de predicción, en lo relativo a las variables implicadas en cada caso, puede observarse en los cuadros de síntesis siguientes:

PREDICTORES DEL FACTOR I

Variables	Curso								
	4.º	6.º	8.º	FP1	FP2	BUP	COU	Univ2º	Univ4º
1. Pronom. Pers.									
2. Voc. García Hoz.									
3. σ palabras			X	X	X	X	X	X	
4. Voc. TV L. Del.			X	X	X	X	X	X	X
5. Pal. por frase			X		X		X	X	X
6. Total puntos						X			
7. Puntos aparte	X	X	X	X	X				
8. Total comas	X		X	X	X	X	X	X	X
9. $X + 2.58 \sigma$	X	X							
10. Puntos coma		X				X	X	X	
11. Nom. propios	X	X	X	X	X	X	X	X	
12. TTR						X	X	X	X

En las ecuaciones de predicción obtenidas a partir de las puntuaciones factoriales del factor I como criterio, se observa la presencia de componentes de los tres factores descriptores de las variables: lexicográfico, fraseológico y de dificultad sintáctica.

A continuación se presenta el mismo gráfico, ahora referido al factor I.

PREDICTORES DEL FACTOR II

Variables \ Curso	4.º	6.º	8.º	FP1	FP2	BUP	COU	Univ2º	Univ4º
1. Pronom. Pers.									
2. Voc. García Hoz.			x						
3. σ palabras									
4. Voc. TV L. Del.									
5. Pal. por frase								x	
6. Total puntos	x								
7. Puntos aparte		x							
8. Total comas									
9. $X + 2.58 \sigma$							x		
10. Puntos coma									
11. Nom. propios							x		
12. TTR	x				x				

Estas predicciones efectuadas a partir de las puntuaciones factoriales calculadas a partir del factor II resultan mucho más desdibujadas.

5. CONCLUSIÓN

La propuesta de ecuaciones de lecturabilidad que hemos realizado pone de manifiesto la posibilidad de avanzar en las cotas de predicción por la vía de la valoración de la comprensión lectora a través de puntuaciones factoriales.

Es de interés igualmente resaltar las posibilidades del indicador "media más 2,58 sigmas, no utilizado hasta ahora en las investigaciones lingüísticas cuantitativas en ningún idioma.

6. BIBLIOGRAFÍA

- ANDERSON, R. C. y FREEBODY, P. (1979): *Vocabulary Knowledge*, Center for the Study of Reading, Urbana, Univ. of Illinois.
- BAYÓN ALVARO, P. (1984): *Validación del test cloze como medida de comprensión lectora*. Memoria de Licenciatura, Universidad de Salamanca.
- BELTRÁN de TENA, R. (1983): El proceso de comprensión lectora. Resultados de una investigación, en *El Sistema Educativo Hoy. Proyecto Cinae*, Buenos Aires, 403-414.
- BUCHANAN, M.A. (1941): *A Graded Spanish Word Book*. Toronto, Univ. of Toronto Press.
- CABERO PÉREZ, M. V. (1991): *Fórmulas de Legibilidad en Lengua Castellana para Los Niveles Educativos de B.U.P. y Enseñanza universitaria*. Tesis doctoral, inédita, Departamento de Didáctica, Universidad de Salamanca.
- CRADWORTH, A. N. (1984): *A Spanish Language Fry-type Readability Procedure. Elementary Level Bilingual Education*. Los Angeles, Paper Series. Evaluation, Dissemination and Assessment Center, California State University.
- FERNÁNDEZ HUERTA, J. (1959): Medidas Sencillas de Lecturabilidad, *Consigna*, 214, 29-32.
- GARCÍA HOZ, V. (1953): *Vocabulario Usual, Vocabulario Común y Vocabulario Fundamental. Determinación y Análisis de sus Factores*. Madrid, C.S.I.C.
- GARCÍA, W. F. (1977): Assessing Readability for Spanish as a Second Language: The Fry Graph And Cloze Procedure. *Dissertation Abstract*, 38, 136a.
- GILLIAN, B.; PEÑA, S. C.; y MOUNTAIN L.: (1980): The Fry Graph Applied to Spanish Readability, *The Reading Teacher*, 3, 1426-430.
- GOICOECHEA TABAR, M. J. (1991): *Competencia lingüística en poblaciones monolingüe y bilingüe Vasco Castellana*. Tesis doctoral, inédita. Departamento de Didáctica, Organización y Métodos de Investigación, Universidad de Salamanca.
- GRUPO DE INVESTIGACIONES CFPJ (1987): Legibilidad y Escritura Telemática, en RICHAUDEAU, F. (ed.), *La legibilidad. Investigaciones actuales*.
- HERRERA GARCÍA, E. (1992): *Evolución del lenguaje escrito en los niños. Un análisis de la estructura superficial de los textos elaborados por alumnos de cuatro niveles de EGB*. Tesis doctoral, inédita, Departamento de Didáctica, Organización y Métodos de Investigación, Universidad de Salamanca.
- HEWITT, G. (1980): A Preliminary Study of Pupils' Reading Difficulties, *Educational Review*, 3, 231 - 244.
- INIC/ Centro de Linguistica da Universidades de Lisboa (1984): *Portugues Fundamental*, Lisboa
- JOHNSTON, P. H. (1989): *La Evaluación de la comprensión lectora*. Madrid, Visor, 31 y ss.
- KANDEL, L. y MOLES, A. A. (1958): Application de l' Indice de Flesch a la Langue Française, en *Cahiers d' Etudes de Radio Television*, 9, 252-275.
- LÓPEZ RODRÍGUEZ, N. (1976): *Indicadores cuantitativos de comprensión lectora en textos narrativos*. Memoria de Licenciatura inédita. Departamento de Didáctica, Universidad de Valencia.
- LÓPEZ RODRÍGUEZ, N. (1981): *Fórmulas de Legibilidad para la Lengua Castellana*. Tesis Doctoral, Departamento de Didáctica, Universidad de Valencia.
- LÓPEZ RODRÍGUEZ, N. (1982): *Cómo valorar textos escolares*. Madrid, Ed. Cincel.
- LORENZO DELGADO, M. (1981): *El Vocabulario Televisivo y su inserción en la enseñanza*, Instituto de Ciencias de la Educación, Universidad de Granada.
- MARTINET, A. (1974): *Elementos de Lingüística General*. Madrid, Gredos.
- MONSALVO DÍEZ, E. (1986-87): Análisis de la prueba cloze en segundo de EGB, *Enseñanza*, 4-5, 155-165.
- MORO BERIHUETE, P. (1991): *Fórmulas de lecturabilidad en lengua castellana para los niveles educativos de E.G.B. y F.P.* Tesis doctoral, inédita, Departamento de Didáctica, Universidad de Salamanca.
- MÜLLER, Ch. (1973): *Estadística Lingüística*. Madrid, Gredos.

- PATTERSON, F.W. (1972): *Cómo Escribir Para Ser Entendido*. El Paso, Casa Bautista de Publicaciones.
- QUINTERO GALLEGO, A. (1986-87): Análisis de la validez del test cloze como prueba de comprensión lectora, *Enseñanza*, 4-5, 149-154.
- RABIN, A. T. (1988): Determining Difficulty Levels of Text Written in Languages other than English", en Zakaluck, B.I. y Samuels, S.j. (ed): *Readability. Its Past, Present & Future*. Ira, Newark, Delaware, 46-76.
- RODRÍGUEZ DIÉGUEZ, J. L. (1983): Evaluación de textos escolares", *Revista de Investigación Educativa*, I(2), 259-279
- RODRÍGUEZ DIÉGUEZ, J. L. (1988): La lecturabilidad del material escrito, en *Cuestiones de Didáctica. Homenaje A. J. Fernández Huerta*, de J.L. Rodríguez Diéguez (ed.). Sociedad Española de Pedagogía. Barcelona, CEAC. 219 - 230.
- RODRÍGUEZ DIÉGUEZ, J. L. (1989): Predicción de la lecturabilidad de textos en castellano: Una propuesta y algunas sugerencias, en *Leer en la Escuela. Nuevas Tendencias en la Enseñanza de la Lectura*. 284 - 310, Madrid. Ed. Pirámide / Fundación Germán Sánchez Ruipérez.
- RODRÍGUEZ DIÉGUEZ, J. L. (1991): Evaluación de la comprensión lectora, en A.PUENTE (ed.): *Comprensión de la lectura y acción docente*. Madrid. Fundación Germán Sánchez Ruipérez y Ed. Pirámide, 301-345.
- RODRÍGUEZ DIÉGUEZ, J. L. y BELTRÁN DE TENA, R. (1985): Comprensión lectora: evaluación cualitativa y cuantitativa de un fenómeno educativo, en *Revista del Instituto de Investigaciones Educativas*, 50, 77-92.
- RODRÍGUEZ DIÉGUEZ, J. L. y LEBRE AIRES, M.L. (1990): Aproveitamento Escolar e Linguagem: Variáveis Individuais, *Revista Portuguesa de Pedagogia*, Año XXIV, 345-366
- SALVADOR MATA, F. (1986): Aspectos didácticos del texto escrito por los alumnos del Ciclo Medio de E.G.B. Estudio evolutivo y diferencial, *Rev. Española de Pedagogía*, 171, 97-115.
- SPAULDING, S. (1951): Two Formulas for Estimating the Reading Difficulty of Spanish, *Educational Research Bulletin*, 30, 117-124
- SPAULDING, S. (1956): A Spanish Readability Formula, *Modern Language Journal*, 40, 433-441
- SPIRO, R. J. (1980): *Schema Theory And Reading Comprehension: New Directions*. Urbana, Univ. of Illinois, Center for the Study of Reading.
- THONIS, E. W. (1976): *Literacy for America's Spanish Speaking Children*. Newark, International Reading Association.
- TINKER, M. A. (1963): *Legibility of Print*. Ames, Iowa, Iowa State University Press.
- ULLMANN, S. (1970): *Semántica. Introducción a la ciencia del significado*. Madrid, Aguilar.
- VARI-CARTIER, P.: (1981): Development and Validation of a New Instrument to Assess the Readability of Spanish Prose, *Modern Language Journal*, 65, 141-148.
- ZIPF, G. K. (1949): *Human Behavior and the Principle of Least Effort*. Cambridge, Addison Wesley Press.