

BASES LINGÜÍSTICAS Y COMPUTACIONALES DEL PROCESAMIENTO DE LA IMPROPIEDAD ESTILÍSTICA Y LA LEGIBILIDAD

JAVIER GÓMEZ GUINOVART
Universidad de Vigo

RESUMEN. *En este trabajo presentamos las bases lingüísticas y computacionales de la verificación automática del estilo, y de la evaluación de la legibilidad o nivel de dificultad de comprensión lectora de los textos. El artículo se divide en cinco apartados: (1) introducción a los objetivos de este trabajo, (2) aproximación teórica al estudio del estilo en lingüística, (3) análisis del concepto de propiedad estilística, (4) presentación de las técnicas de procesamiento de la impropiedad estilística basadas en modelos textuales, y (5) descripción de los métodos estadísticos de análisis lingüístico-cuantitativo de la legibilidad y de su aplicación al análisis de textos en castellano.*

PALABRAS CLAVE. *Verificación automática del estilo, estilística lingüística, modelos textuales, estilometría, análisis de la legibilidad, lingüística estadística, lingüística cuantitativa.*

ABSTRACT. *This paper presents the linguistic and computational foundations of automatic style checking, and of the evaluation of the readability or reading comprehension level of texts. The paper is divided into five sections: (1) an introduction to the objectives, (2) a theoretical approach to the study of style in linguistics, (3) an analysis of the concept of stylistic propriety, (4) a presentation of the techniques in stylistic improprieties processing based on textual models, and (5) a description of the statistical methods of quantitative linguistics for the evaluation of readability, and of its application to Spanish texts.*

KEYWORDS. *Automatic style checking, linguistic stylistics, textual models, stylometry, evaluation of readability, statistical linguistics, quantitative linguistics.*

1. INTRODUCCIÓN

La verificación automática de la corrección lingüística constituye una de las aplicaciones informáticas de mayor incidencia en la calidad de los resultados del procesamiento de textos. En esta categoría de aplicaciones, las herramientas más utilizadas y mejor consideradas son los correctores ortográficos. A su lado, los

programas de verificación gramatical y estilística (muchos de ellos, aún en fase de desarrollo) poseen un nivel de aceptación mucho menor y una eficacia en ocasiones cuestionable. En este artículo, trataremos de presentar de forma pormenorizada los fundamentos lingüísticos y computacionales de la verificación automática del estilo, y de la evaluación de la legibilidad o nivel de dificultad de lectura de los textos.

Aunque la mayoría de las personas aceptarían definir sucintamente el estilo como la “manera de escribir o de hablar”, la gran variedad de acepciones con que se emplea este vocablo aconseja aproximarse a él con cierta cautela y vasta amplitud de miras. Por esta razón, dedicaremos el primer apartado de este trabajo a caracterizar algunas de las principales aproximaciones teóricas al estudio del estilo, desde una perspectiva eminentemente lingüística y centrada en la estilística de la variación. A continuación, analizaremos el ámbito de la propiedad del estilo, y comprobaremos las relaciones que se establecen entre este concepto, el de aceptabilidad gramatical y el de variedad lingüístico-estilística. Finalmente, examinaremos diversas estrategias informáticas utilizadas en el procesamiento automático de la impropiedad estilística, como la aplicación de modelos y tipologías textuales, la gradación de la impropiedad estilística, y el análisis lingüístico-cuantitativo de la legibilidad.

2. ESTILÍSTICA Y VARIACIÓN LINGÜÍSTICA

Dentro de las diversas concepciones de la investigación estilística, pueden distinguirse dos vertientes principales en función de su ámbito de estudio: la estilística literaria, que dirige sus intereses hacia el lenguaje literario; y la estilística lingüística, cuyo objetivo es el lenguaje en todos sus registros o variedades.

En líneas generales, la estilística literaria recurre al análisis lingüístico de los textos como fundamento del análisis crítico, y concibe el estilo como un desvío de la norma lingüística estándar y como una elección consciente, por parte del autor, entre las diferentes posibilidades expresivas que le ofrece la lengua. Esta orientación de la estilística –denominada a veces “crítica estilística” y representada por autores como Karl Vossler, Leo Spitzer o Dámaso Alonso– se centra en el análisis lingüístico-estético de la obra literaria, y suele situarse en un terreno intermedio entre la lingüística y la crítica literaria. Así, uno de los investigadores más prolíficos en el campo de la estilística literaria, Henry Widdowson, define la estilística como “el estudio del discurso literario desde una perspectiva lingüística”. Según este autor, la estilística es una manera de vincular la crítica literaria con la lingüística:

By ‘stylistics’ I mean the study of literary discourse from a linguistic orientation and I shall take the view that what distinguishes stylistics from literary criticism on the one hand and linguistics on the other is that it is essentially a means of linking the two and has (as yet at least) no autonomous

domain of its own [...]. Given a piece of literature, a poem for example, the linguist will be interested in finding out how it exemplifies the language system, and if it contains curiosities of usage how these curiosities might be accounted for in grammatical terms [...]. The literary critic, however, takes interpretation as his aim [...]. The purpose of stylistics is to link the two approaches by extending the linguist's literary intuitions and the critic's linguistic observations and making their relationship explicit. (Widdowson 1975: 3-6)

Por otro lado, la estilística lingüística contemporánea suele concebir el estilo como el conjunto objetivo de características formales ofrecidas por un texto como resultado de la variación lingüística o de las intenciones de su autor. Dentro de este marco general, existen diferentes concepciones sobre cuáles son los rasgos que caracterizan lingüísticamente el estilo de los textos, y cuál es la situación de la estilística con respecto a las otras ciencias del lenguaje.

Por ejemplo, en el ámbito teórico del estructuralismo, Coseriu (1981) define la estilística como una subdisciplina de la lingüística descriptiva centrada en el estudio de la variación interna de las lenguas producida por diferencias diafásicas. Recogiendo la terminología empleada por Coseriu, la variación intralingüística en el eje sincrónico está motivada por tres tipos de diferencias: diatópicas o diferencias en el espacio geográfico, diastráticas o diferencias entre los distintos niveles socioculturales de la comunidad hablante, y diafásicas o diferencias relacionadas con la circunstancia concreta del acto lingüístico.

De acuerdo con este autor, estos tres tipos de diferencias producen tres tipos de variedades intralingüísticas: las variedades sintópicas o dialectos, las variedades sinstráticas o niveles de lengua (lenguaje culto, lenguaje de la clase media, lenguaje popular, etc.), y las variedades sinfásicas o estilos de lengua (lenguaje familiar, lenguaje solemne, etc.). Dentro del grupo de las variedades sinfásicas o estilos de lengua, Coseriu incluiría las variaciones propias de un género (el lenguaje de las mujeres) o de una generación (el lenguaje de los adultos); y las variaciones características de los grupos sociales (el lenguaje de los estudiantes) o profesionales (el lenguaje de los filatélicos), siempre que “puedan distinguirse en el mismo nivel sociocultural (o independiente de los niveles)”. Coseriu denomina “registros idiomáticos” a los tipos muy generales de estilos de lengua correspondientes a aspectos amplios de la vida y de la cultura (por ejemplo, lengua hablada, lengua escrita o lengua literaria). Por último, dentro de este marco teórico, un sistema lingüístico unitario desde los tres puntos de vista (diatópico, diastrático y diafásico) recibe el nombre de “lengua funcional”. Por ejemplo, el español culto actual hablado en Quito en las situaciones formales por la clase alta sería una lengua funcional o, dicho de otro modo, una “variedad lingüístico-estilística”.

A partir de esta caracterización de la variación lingüística, las subdisciplinas de la lingüística que se encargarían de estudiarla serían, de acuerdo con este autor: la dialectología, que estudiaría las variedades sintópicas o dialectos; la sociolingüística,

enfocada hacia las variedades sinstráticas o niveles de lengua; y la estilística, que se centraría en las variedades sinfásicas o estilos de lengua

Otro punto de vista algo distinto, procedente del funcionalismo y la sociolingüística, distingue dos categorías o dimensiones básicas de la variación lingüística: la variación en función del hablante y la variación en función del uso lingüístico. Halliday y Hasan (1976), por ejemplo, denominan a estas dos clases de variación, respectivamente, “dialectos” y “registros”; y Gregory (1967), “variantes dialectales” y “variantes diatópicas”. Siguiendo a este autor, las primeras serían aquellas variaciones debidas a características razonablemente permanentes del hablante en cualquier situación lingüística, como la individualidad lingüística (variación idiolectal), la procedencia temporal (variación dialectal temporal), la procedencia geográfica (variación dialectal geográfica) o la procedencia social (variación dialectal social). Las variantes diatópicas serían, en cambio, el reflejo lingüístico de las características del distinto uso del lenguaje que haría el hablante en diferentes situaciones comunicativas según la finalidad del enunciado (o campo del discurso), según el medio empleado (o modo del discurso), según la relación personal con los interlocutores (o tenor personal), y según la relación funcional con los interlocutores (o tenor funcional).

En esta perspectiva, la estilística lingüística estudiaría los rasgos lingüísticos asociados con las variedades diatópicas o registros (en el sentido de Halliday) y relacionados con el campo, modo y tenor específicos del discurso, si bien hay autores que proponen extender su campo de estudio a la totalidad de los factores de variación.

Así, Crystal y Davy (1969) proponen una estilística orientada al análisis textual de los rasgos lingüísticos que caracterizan las variedades principales de una lengua. De acuerdo con Crystal y Davy, estos rasgos lingüísticos pueden asignarse a ocho dimensiones estilísticas aproximadamente coincidentes con los ocho factores de variación distinguidos por Gregory, y todos los textos podrían contener información respecto a estas ocho dimensiones estilísticas. Dentro de cada dimensión estilística podrían señalarse diversas categorías situacionales, susceptibles de ser definidas en función de sus rasgos lingüísticos distintivos. Por ejemplo, dentro de la dimensión estilística del tenor del discurso, podrían establecerse diversos grados de formalidad –diversas categorías situacionales–, como rígido, formal, normal, informal e íntimo; y dentro de la dimensión de variación por el campo o dominio del discurso, podría diferenciarse el lenguaje de los anuncios, el científico, el jurídico o el del culto religioso.

Los rasgos lingüísticos distintivos de una categoría situacional perteneciente a una dimensión estilística se denominan, en la terminología de Crystal y Davy, “rasgos estilísticos” o “indicadores de variedad”. Una variedad lingüístico-estilística se describiría como una combinación concreta de rasgos lingüísticos pertenecientes a diversas categorías situacionales (por ejemplo, la variedad “sermón” podría describirse como un conjunto de rasgos lingüísticos típicos de las categorías oral,

formal, culto religioso, monólogo, etc.), y coincidiría total o parcialmente con lo que otros autores denominan “sublenguajes estilísticos”, “géneros”, “registros” o “lenguas funcionales”.

En suma, la estilística de la variación sistematizada por Crystal y Davy (1969) establece como objetivo central de esta disciplina la descripción del conjunto distintivo de rasgos lingüísticos que caracterizan cada una de las variedades lingüístico-estilísticas de una lengua. La investigación partiría de una tipología de textos y dedicaría todos sus esfuerzos a la descripción y catalogación de los rasgos lingüísticos definitorios de cada clase.

Sin embargo, el modelo de Crystal y Davy no ofrece ningún método objetivo para establecer el repertorio completo de géneros que tipifican los textos de una lengua, sino que supone una tipología textual no explícita, instituida *a priori* por medios no muy bien definidos, entre los que se vislumbran las intuiciones del hablante, las situaciones en que se utilizan los textos y su configuración externa. A tal respecto, Biber (1988) –y, más recientemente, Kessler et al. (1997)– propone realizar investigaciones empíricas, en corpus amplios y representativos de la lengua, que permitan determinar mediante métodos estadísticos una tipología objetiva de los tipos de textos de una lengua, considerando exclusivamente sus rasgos lingüísticos.

Los resultados combinados de los estudios sobre tipología textual lingüístico-cuantitativa y sobre estilística de la variación, aplicados convenientemente, podrían constituir un factor decisivo para mejorar la eficacia de las aplicaciones informáticas más difundidas de verificación del estilo, ya que –como veremos a continuación– parte de estas aplicaciones basan su estrategia de verificación en la adscripción del texto analizado a una categoría textual y en la comparación de los rasgos lingüísticos característicos de esta categoría con los rasgos detectados en el texto.

3. IMPROPIEDAD Y ACEPTABILIDAD

A partir de las consideraciones teóricas expuestas por la estilística de la variación, asumiremos que la propiedad estilística de un texto se basa en la afinidad de su estilo con la norma estilística del género al que se adscribe dicho texto, y asumiremos también que la norma estilística de un género determinado puede definirse a partir de las construcciones lingüísticas de uso más común en tal género. Con estos presupuestos, resulta obvio señalar que cualquier intento de determinar el grado de propiedad (o de impropiiedad) del estilo de un texto requiere establecer previamente cuáles son las variantes lingüístico-estilísticas o géneros de una lengua, y cuáles son los rasgos lingüísticos que caracterizan cada una de las variantes postuladas (es decir, cuáles son los rasgos que definen la norma estilística de cada género definido).

Estudios como los de Biber demuestran la trascendencia de realizar investigaciones empíricas que permitan determinar mediante métodos estadísticos

(como el análisis factorial y el análisis de conglomerados) una tipología objetiva de los géneros de una lengua, en función de los rasgos lingüísticos sintácticos y léxicos que comparten con mayor frecuencia. De acuerdo con el método cuantitativo aplicado por Biber, las características lingüísticas mayoritarias y estadísticamente centrales del conglomerado de textos agrupados en un género constituirían la configuración lingüística prototípica o norma estilística de dicho género. Adoptando esta perspectiva, un programa informático de verificación estilística automática podría detectar las impropiedades del estilo de un texto, con respecto a una determinada norma estilística, comparando los rasgos lingüísticos del texto analizado con los rasgos lingüísticos prototípicos que caracterizan la norma estilística del género al que se adscribe dicho texto.

En consecuencia, un mismo rasgo lingüístico puede considerarse propio o impropio del estilo de un texto, según la variedad lingüístico-estilística a la que pertenezca el texto que lo contiene. Por ejemplo, basándonos en los datos obtenidos por Biber (1989) a partir del análisis estadístico de un corpus amplio de textos orales y escritos pertenecientes a una amplia gama de géneros, la presencia de la contracción *don't* en un texto en inglés se consideraría estilísticamente apropiada en una carta personal informal o en una conversación, pero no en una póliza de seguros o en un documento oficial, donde lo más apropiado estilísticamente (o sea, lo estadísticamente más frecuente y central) sería la forma *do not* no contracta.

Así mismo, podemos establecer una correspondencia entre el concepto de impropiedad estilística y el concepto de aceptabilidad, introducido por Chomsky (1965) como medida de la actuación lingüística: una oración es aceptable si no resulta extraña a causa de ciertos factores estilísticos, ni hace falta papel y lápiz para entenderla debido a su complejidad sintáctica. Ejemplo de secuencias gramaticales con problemas de aceptabilidad son:

- a. *Los antibióticos (substancias producidas por microorganismos que a bajas concentraciones inhiben o matan otros microorganismos), usados hoy contra la tuberculosis –auténtica plaga en otros tiempos– y contra muchas otras enfermedades infecciosas, salvan millones de vidas*
- b. *La rata que el gato cazó murió*
- c. *Yo y tú iremos mañana al cine*
- d. *La rata que el gato que el perro cazó comió murió*

Es evidente que los enunciados (a) y (b) ofrecen unas dificultades intrínsecas de comprensión derivadas de la complejidad de sus respectivas estructuras sintácticas. Sin embargo, como observan Hernanz y Brucart, la aceptabilidad global de este tipo de enunciados puede depender en gran medida de determinadas condiciones de su situación comunicativa (por ejemplo, del tiempo de procesamiento y del canal de comunicación empleado):

Uno de los objetivos de la teoría de la actuación será medir el grado de complejidad a partir del cual un enunciado se convierte en inaceptable. Es probable que el umbral que delimite tal medida varíe según el soporte físico del mensaje. Una oración que haga uso reiterado de la incrustación puede resultar incomprensible para el oyente cuando se emite oralmente. Sin embargo, si se le suministra el mismo enunciado por escrito y se le da el tiempo necesario, la misma persona será capaz de interpretarlo de manera correcta. (Hernanz y Brucart 1987: 16).

Por otra parte, los problemas de aceptabilidad de (c) no dependen de factores sintácticos, sino de factores estilísticos relacionados con ciertas convenciones sociales de cortesía en el lenguaje. Como señalan los mismos autores, el enunciado de (c) “altera una convención social de urbanidad según la cual el pronombre referido al hablante, cuando concurre con cualquier otro nombre o pronombre, debe ir colocado en última posición”.

Desde el punto de vista de la estilística de la variación, la mayor parte de los factores capaces de determinar la aceptabilidad de un enunciado –como el tiempo de procesamiento, el canal de comunicación o el respeto de una determinada convención social– pueden explicarse por las características lingüístico-estilísticas del tipo de texto (de la variedad, del género) al que pertenece tal enunciado. En consecuencia, es probable que los enunciados de (a) y (b) sean absolutamente comprensibles y relativamente aceptables en las variedades que utilizan el canal escrito y que suelen implicar una lectura pausada, como en los textos literarios cultos; y, al mismo tiempo, sean prácticamente inaceptables en un género escrito de lectura rápida, como el de los textos informativos periodísticos, o en un género oral, como el de las soflamas políticas espontáneas. Del mismo modo, es probable que el enunciado de (c) sea inaceptable en aquellos registros lingüísticos formales que suponen un respeto estricto de las convenciones sociales, por ejemplo, en la defensa oral de un acusado ante un tribunal de justicia por parte de su abogado; y, en cambio, el mismo enunciado pueda ser relativamente aceptable y bastante menos chocante dentro de una variedad lingüístico-estilística donde la adhesión a las normas de urbanidad sea mucho menos rigurosa, por ejemplo, en una conversación telefónica íntima.

Por tanto, admitiendo que la aceptabilidad de un enunciado depende en gran medida del tipo de texto en que aparece, podemos establecer una cierta equivalencia funcional entre el concepto de aceptabilidad y el concepto de propiedad estilística: un mismo rasgo lingüístico –como la incrustación de un constituyente extenso y complejo en (a), la presencia de una autoincrustación de relativo en (b) o el incumplimiento de una convención social de urbanidad en el lenguaje en (c)– puede considerarse propio o impropio del estilo de un texto, aceptable o inaceptable, según la variedad lingüístico-estilística a la que pertenezca el texto que lo contiene. Esta correspondencia entre los conceptos de aceptabilidad y de propiedad estilística no implica necesariamente que no existan enunciados inaceptables (y, por tanto,

impropios) en prácticamente cualquier género de textos, como lo son posiblemente las construcciones similares a (d), con varios niveles de autoincrustación de relativo.

4. MODELOS TEXTUALES PARA LA VERIFICACIÓN AUTOMÁTICA DEL ESTILO

Una de las técnicas más utilizadas por los sistemas de verificación automática del estilo, la asignación del texto analizado a una determinada variedad lingüístico-estilística, se basa precisamente en la constatación de que la aceptabilidad de un enunciado depende en gran medida del tipo de texto en que aparece. Antes de que el programa lleve a cabo la verificación estilística de un documento, la persona usuaria del sistema debe especificar a qué categoría pertenece el texto en cuestión. Así, el programa puede efectuar la verificación del estilo comparando las características del documento analizado con los rasgos lingüístico-estilísticos que el sistema establece para la categoría textual seleccionada.

Por ejemplo, el programa de verificación estilística de textos en inglés Ruskin (McGowan 1992) precisa que se le proporcionen ciertas estimaciones sobre el documento que se pretende revisar, con respecto a su audiencia (edad, años de escolarización, grado de familiaridad con el tema, etc.), al propósito del texto (didáctico, argumentativo, divulgativo, etc.), al tema desarrollado (nivel de complejidad, uso de terminología específica, etc.), al uso que se le va a dar al documento (consumo interno, obra de referencia, número de lectores, etc.) y a la persona autora del texto (nivel de dominio del tema, facilidad de escritura, etc.).

Utilizando las estimaciones proporcionadas para las variables contextuales del documento y un conjunto de reglas condicionales del tipo “si-entonces”, Ruskin construye un modelo o norma estilística al que idealmente debería ajustarse el documento analizado. Dos ejemplos concretos de estas reglas condicionales del “buen estilo”, utilizadas por Ruskin para la evaluación de textos en inglés, serían los siguientes:

- a. si *audiencia es universitaria*
entonces *media_de_palabras_por_oración* debe ser igual o menor de 12,3
- b. si *audiencia es desconocida*
entonces *máximo_de_palabras_por_oración* debe ser 28

En caso de que Ruskin detectara, en el documento analizado, cualquier característica opuesta a las especificaciones “ideales” codificadas en las reglas condicionales correspondientes a su categoría estilística, el sistema señalaría una impropiedad de estilo e indicaría su posible solución.

En otros programas de verificación lingüística, como Grammatik, se puede seleccionar la norma con la que se va a contrastar el documento a partir de un número de modelos estilísticos predefinidos. Cada uno de estos modelos textuales está definido en virtud de un conjunto de rasgos lingüístico-estilísticos determinados, como la longitud máxima de una oración, la presencia o ausencia de términos y construcciones coloquiales, o el número máximo de sintagmas preposicionales consecutivos. Así mismo, las personas usuarias de Grammatik pueden crear sus propios modelos estilísticos, asignando los valores deseados a las características lingüísticas propuestas por el programa. Por ejemplo, los diez modelos estilísticos predefinidos en Grammatik V (WordPerfect Corp., 1993) para el inglés británico, distribuido en la versión 6 para DOS del procesador de textos WordPerfect, son:

1. general
2. cartas comerciales
3. notas internas
4. informes financieros y textos académicos
5. textos técnicos
6. manuales de instrucciones
7. propuestas y proyectos
8. textos periodísticos
9. textos publicitarios
10. textos narrativos

El hecho de asignar un texto a uno de estos diez modelos estilísticos determina tanto el conjunto de rasgos lingüístico-estilísticos del texto que van a ser revisados, como los umbrales de impropiedad de algunos de estos rasgos. Como ilustración de esta técnica de verificación, consideremos los siguientes rasgos estilísticos del inglés comprobados por Grammatik V, dicotómicos los tres primeros, y cuantificables los restantes:

- A. Presencia de términos y construcciones coloquiales
- B. Posición de los adverbios temporales de frecuencia indefinida en posición medial
- C. Uso de la segunda persona gramatical
- D. Número máximo de nombres consecutivos
- E. Número máximo de sintagmas preposicionales consecutivos
- F. Número máximo de verbos en forma pasiva por cada diez construcciones verbales
- G. Número máximo de palabras por oración.

La siguiente tabla resume los valores asignados a cada uno de estos siete rasgos estilísticos en los diez modelos textuales predefinidos por Grammatik V, con los nombres descriptivos de los distintos estilos en el eje vertical, los rasgos lingüísticos considerados en el eje horizontal, y los valores asignados a cada rasgo en cada estilo en las intersecciones correspondientes:

	A	B	C	D	E	F	G
<i>General</i>	s.	s	n	4	4	1	30
<i>Cartas comerciales</i>	s	s	n	4	4	1	35
<i>Notas internas</i>	s	s	n	4	4	1	25
<i>Informes financieros y textos académicos</i>	s	s	s	5	5	2	40
<i>Textos técnicos</i>	s	s	s	6	5	2	40
<i>Manuales de instrucciones</i>	s	s	n	4	4	1	30
<i>Proyectos y propuestas</i>	s	s	s	5	5	3	35
<i>Textos periodísticos</i>	n	s	n	5	4	2	30
<i>Textos publicitarios</i>	n	n	n	5	4	2	25
<i>Textos narrativos</i>	n	n	n	5	4	3	40

Evidentemente, la efectividad de la verificación estilística por modelización del texto será proporcional al acierto de la tipología textual utilizada y al grado de exactitud con el que estén establecidos los distintos parámetros de evaluación de sus características lingüísticas. De ahí la importancia fundamental que poseen para el desarrollo de esta técnica los avances realizados en los campos de la tipología textual lingüístico-cuantitativa y de la estilística de la variación.

4.1. Grados y umbrales de impropiedad

Como acabamos de mencionar, la utilización de tipologías textuales para la verificación automática del estilo suele complementarse con la posibilidad de modificar las opciones concretas de verificación para cada rasgo estilístico. En el caso de los rasgos considerados cuantificables, como el número máximo de palabras permitido por oración, la gradación de la impropiedad estilística permite que un rasgo lingüístico determinado sea considerado una impropiedad del estilo cuando supere un determinado umbral. La dificultad surge cuando el rasgo graduable que se pretende revisar es algo tan indeterminado como la distancia excesiva entre el sujeto y el verbo –cuestión que, en términos formales, implica definir necesariamente un valor cuantificable para el concepto de distancia excesiva–. Veamos, a modo de ejemplo,

uno de los enfoques adoptados para la solución de este problema, en el marco concreto del programa de verificación para el inglés del procesador de textos Critique (Jensen *et al.* 1993), desarrollado por IBM.

La estrategia, propuesta por Heidorn *et al.* (1982), consiste en especificar una función CRIT(DSV) –distancia crítica entre el sujeto y el verbo– que combine los factores sintácticos que inciden en la percepción del alejamiento entre el verbo y su sujeto, a saber, la posición del núcleo del sujeto –POS(SUJ)–, la posición del verbo correspondiente –POS(V)– y el número de sustantivos presentes entre el núcleo del sujeto y el verbo –NUM(N)–. En la función CRIT(DSV) que estos autores sugieren se incluye un factor de ponderación p para NUM(N), mientras que la utilización del valor absoluto de POS(V)–POS(SUJ) permite manejar los casos de inversión sujeto-verbo:

$$\text{CRIT(DSV)} = |\text{POS(V)} - \text{POS(SUJ)}| \times (1 + p \times \text{NUM(N)})$$

Una vez formulada la función, el siguiente paso es establecer un valor para el factor de ponderación p y un valor crítico (un umbral de impropiedad) para CRIT(DSV), de tal manera que cuando el resultado de la función supere este umbral, el sistema señale una impropiedad estilística. Cuanto menor sea el umbral, menor será la sensibilidad de la función para detectar esta impropiedad. Por ejemplo, dado un valor de 0,2 para el factor de ponderación p y un valor de 10,1 para el umbral de impropiedad de CRIT(DSV); y dada una oración con el núcleo del sujeto como segunda palabra –POS(SUJ)=2– y sin sustantivos entre el sujeto y el verbo –NUM(N)=0–, esta función detectará una impropiedad cada vez que el verbo se encuentre en la decimotercera palabra –POS(V)=13– o más adelante, ya que el valor 11 resultante de la función CRIT(DSV) superará el umbral establecido:

$$\text{CRIT(DSV)} = (13 - 2) \times (1 + (0,2 \times 0)) = 11 > 10,1$$

En la misma circunstancia, pero con dos sustantivos entre sujeto y verbo, bastaría con que el verbo se hallara en la décima palabra:

$$\text{CRIT(DSV)} = (10 - 2) \times (1 + (0,2 \times 2)) = 11,2 > 10,1$$

De este modo, las personas usuarias del procesador de textos Critique podrían adaptar su sistema de verificación del estilo a sus necesidades específicas, ajustando a su medida tanto el factor de ponderación de NUM(N), como el umbral de impropiedad de la función CRIT(DSV).

5. ESTILÍSTICA Y ESTADÍSTICA

Una de las características más sorprendentes del uso que las personas hablantes hacen de su lengua es la regularidad estadística con la que parecen controlar muchos aspectos de su sintaxis, su léxico, su pronunciación y su escritura. En el uso contemporáneo del léxico castellano, por ejemplo, según observan Patterson y Urrutibéheity (1975), aunque un diccionario estándar contiene unas 100.000 entradas léxicas, con las 100 palabras usadas con mayor frecuencia en esta lengua se puede componer más del 30% del material léxico de cualquier texto en español; con las 1.000 palabras más usadas, más del 50%; y con las 5.000 más usadas, más del 90%. Por otra parte, en el nivel de la pronunciación, alrededor de un 50% de las sílabas emitidas en castellano actual son del tipo CV (es decir, consonante seguida de vocal, como las dos sílabas de *ca-sa*); un 20%, del tipo CVC (como las dos de *por-tal*); cerca de un 10% son del tipo V; el mismo porcentaje le corresponde al tipo VC; y el 10% restante se lo reparten los demás tipos silábicos (Guerra 1983; Gómez Guinovart 1993). En cuanto a la frecuencia de uso de los fonemas, tanto las vocales como las consonantes del castellano aparecen en los textos en una proporción similar. Entre las vocales, los más usados son los fonemas /a/ y /e/, únicos fonemas que superarían la frontera del 10%; entre los consonantes, los fonemas /s/ (cerca del 8%) y /l/ (alrededor del 5%) (Rojo 1991).

Es importante señalar que las personas hablantes no aplican estos porcentajes de forma consciente y, sin embargo, aparecen las mismas regularidades en cualquier muestra lingüística suficientemente extensa. Hay quien llega a afirmar, incluso, que existen regularidades estadísticas comunes a todas las lenguas. Así, la “ley de Zipf” (así denominada en honor de su autor, el filólogo americano George Kingsley Zipf) establece la existencia de una relación matemática constante entre el número de apariciones de una palabra en un texto y su posición en una lista de las palabras de ese texto ordenadas por su frecuencia. Zipf también postuló la existencia de una relación inversa entre la longitud de una palabra y la frecuencia con que aparece en los textos. Según este autor, ambas regularidades se darían al margen de la lengua, del autor o del tema, por lo que constituirían una característica universal de las lenguas humanas.

Otra de las facetas del lenguaje donde se ha detectado cierta estabilidad en el porcentaje de uso de determinados rasgos lingüísticos es la variación intralingüística y, muy particularmente, la variación dialectal temporal (o diacrónica) y la variación idiolectal. Basándose en estas regularidades, la estilometría (o estilística estadística) trata de contribuir a determinar la datación cronológica o la autoría de las obras literarias de las que se desconoce su adscripción temporal exacta, o cuya autoría es dudosa o desconocida (Holmes 1994, 1998). En cualquiera de estas dos modalidades, la característica lingüística analizada con mayor frecuencia y provecho en los estudios estilométricos es el material léxico de los textos, y la hipótesis básica de trabajo es que cada persona posee una huella estilística, es decir, unos rasgos de estilo característicos

e inconscientes, que son independientes de su voluntad artística y se manifiestan en toda su obra.

Por ejemplo, supongamos que se sabe que un texto T es obra o de X o de Y, pero no se sabe de quién con total certeza. Si el texto T se hallara en soporte digital, podríamos descubrir sin ninguna dificultad cual es el porcentaje P de apariciones de la palabra *hasta* con respecto al número total de palabras del texto. Se trataría, a continuación, de averiguar el valor del mismo dato en las obras completas de X y en las obras completas de Y, y de comprobar cual de los dos se encuentra más cercano de P. Con un conjunto bien establecido de comparaciones similares, podríamos llegar a conclusiones estadísticamente válidas sobre el grado de probabilidad de cada una de los dos hipótesis de autoría.

La determinación de autorías controvertidas mediante técnicas estadísticas posee una gran importancia también para la investigación criminal (McMenamin 1993). La estilística forense aplica métodos estilométricos, por ejemplo, para determinar la autenticidad de la carta de despedida de una persona que en apariencia se acaba de suicidar, pero que se sospecha que la pueden haber asesinado y que la carta es falsa.

5.1. *Análisis de la legibilidad*

Otra de las aplicaciones de la estilometría consiste en la evaluación del nivel de “legibilidad”, definido como el nivel de dificultad de comprensión del sentido de un texto determinado por ciertos factores lingüístico-estilísticos cuantificables, como la longitud de las oraciones, la longitud de las palabras, la rareza de las palabras (es decir, el porcentaje de palabras del texto no incluidas en una lista determinada de palabras “normales” de la lengua), la cantidad de preposiciones en una frase, etc.

Los estudios de la legibilidad (iniciados en Estados Unidos durante la década de 1920) se basan en las regularidades estadísticas detectadas en ciertas características lingüísticas de los textos, en función de su grado de facilidad o dificultad de lectura. Uno de sus objetivos consiste en la elaboración de “fórmulas de legibilidad”, contrastadas empíricamente mediante métodos estadísticos, que sirvan para predecir parámetros estimativos objetivos del grado de dificultad de un texto en función de su estilo. Esta perspectiva del estudio de la legibilidad de los textos suele distinguir tres factores, por lo menos, que inciden en su nivel de comprensión lectora (Klare, 1963): la legibilidad gráfica, relacionada con el formato externo del texto; la legibilidad semántica, relacionada con su contenido; y la legibilidad lingüístico-estilística, relacionada con su estilo.

En el ámbito español, Núñez Ladevéze (1979) utiliza los términos “legibilidad” y “comprensibilidad” para referirse, respectivamente a la “legibilidad gráfica” y a la “legibilidad lingüístico-estilística”, mientras que Rodríguez Diéguez (1989) emplea

los términos “legibilidad” y “lecturabilidad” –neologismo ausente del diccionario académico, cuya primera documentación la hallamos en Fernández Huerta (1959)– para aludir, respectivamente, a estos dos conceptos. Hacemos constar aquí de paso que, en lo que queda del artículo, utilizaremos el término “legibilidad” para referirnos de manera abreviada a la “legibilidad lingüístico-estilística”, empleando este último término sólo cuando su forma reducida pueda dar ocasión a interpretaciones equívocas.

Las fórmulas de legibilidad constituyen el resultado más popular de los estudios sobre la legibilidad lingüístico-estilística de los textos. Su enorme difusión, especialmente entre las editoriales y agencias de noticias norteamericanas, y la sencillez de su aplicación mecánica suelen ocultar, en muchos casos, la solidez de los fundamentos estadísticos sobre los que se sustentan. El procedimiento clásico para su elaboración suele seguir los cuatro pasos siguientes (Klare y Buck 1954):

1. establecer el conjunto de rasgos lingüísticos objetivables que supuestamente inciden en el nivel de legibilidad de un texto
2. obtener mediante encuestas los índices de dificultad de los textos de un corpus, elaborado como modelo de la variedad lingüística a examen
3. calcular la correlación estadística existente entre los rasgos estilísticos preestablecidos y los índices de dificultad obtenidos empíricamente
4. elaborar una fórmula de legibilidad, a modo de ecuación, con los parámetros estilísticos de mayor capacidad predictiva y menor grado de correlación mutua

Las fórmulas de legibilidad de mayor aceptación utilizan un conjunto de elementos estilísticos muy reducido que, además, no son difíciles de computar, como la media de sílabas por palabra o la media de palabras por oración. Otras fórmulas exigen recursos computacionales más complejos, como la utilización de listas de frecuencias léxicas (para calcular el número de palabras con una determinada frecuencia de uso), diccionarios categorizados (para determinar la cantidad de vocablos pertenecientes a una determinada categoría gramatical) o, incluso, analizadores sintácticos (para comprobar, por ejemplo, las veces que aparece un determinado tipo de construcción sintáctica en el texto).

Sin duda, la fórmula de legibilidad más utilizada para el inglés es la propuesta por Flesh (1948), que calcula el índice de legibilidad (IL) de un texto a partir de la media de sílabas por cada 100 palabras (SP) y la media de palabras por oración (PO), según la siguiente ecuación:

$$IL = 206,835 - 0,846 SP - 1,015 PO$$

Los resultados deben interpretarse mediante la siguiente tabla de equivalencias, que se acompaña de los patrones estilísticos regulares respecto a los valores de SP y PO en cada grado de dificultad:

<i>IL</i>	<i>Grado de dificultad</i>	<i>SP</i>	<i>PO</i>
0-30	Muy difícil	192 o más	29 o más
30-50	Difícil	167	25
50-60	Más bien difícil	155	21
60-70	Estándar	147	17
70-80	Más bien fácil	139	14
80-90	Fácil	131	11
90-100	Muy fácil	123 o menos	8 o menos

Otras dos fórmulas de legibilidad empleadas en esta lengua son las elaboradas por Farr et al. (1951) y por Gunning (1952). La fórmula de Farr, cuyo resultado debe interpretarse mediante la misma tabla de valores que la fórmula de Flesh, calcula el índice de legibilidad (IL) de un texto a partir de la media de monosílabos por 100 palabras (MP) y la media de palabras por oración (PO), realizando el siguiente cómputo:

$$IL = 1,599 MP - 1,015 PO - 31,517$$

Por otra parte, la fórmula de legibilidad por Gunning (1952) calcula el índice de legibilidad (IL) de un texto a partir del porcentaje de palabras de tres o más sílabas en el texto (TT) y la media de palabras por oración (PO), del siguiente modo:

$$IL = 0,4 (TT + PO)$$

El resultado obtenido al aplicar la fórmula de Gunning a un texto representa los años de escolarización necesarios para entender sin dificultades dicho texto.

5.2. Fórmulas de legibilidad para el castellano

En general, la aplicación de las fórmulas de legibilidad del inglés a textos escritos en lengua castellana produce unos resultados del todo inadecuados, clasificando como ilegibles cualquier tipo de textos de legibilidad claramente estándar. La causa de este desajuste radica en que los parámetros estilísticos empleados por estas fórmulas (como la media de sílabas por palabra, la media de palabras por oración, el porcentaje de

monosílabos en el texto, o el porcentaje de palabras de tres sílabas o más) han sido calibrados de acuerdo con las pautas de la lengua inglesa. Como se puede observar en la tabla anterior, según los datos obtenidos por Flesh, un texto de legibilidad estándar en inglés posee una media estimada típica de 1,47 sílabas por palabra y de 17 palabras por oración. En castellano, sin embargo, según los datos ofrecidos por Patterson y Urrutibéheity (1975), la media estimada de sílabas por palabra de un texto representativo es de 1,66 sílabas por palabra, un parámetro estilístico de valor más elevado que el correspondiente para el inglés. En cuanto a la media de palabras por oración, y a pesar de la ausencia de estudios concluyentes sobre la longitud oracional en un corpus representativo de la lengua castellana, coincidimos con Vázquez-Ayora (1977) en la opinión de que “todos los que han practicado la traducción anglo-española tienen la experiencia de que la versión española tiende a ser mucho más larga que en el caso inverso”. Así mismo, el manual de estilo periodístico de El País (1994) recomienda, como longitud oracional máxima aconsejable, no más de 20 palabras y, el de La Vanguardia (1986), “unas 20 palabras sin contar artículos y conjunciones”.

A pesar de todo, resulta bastante común el error de aplicar sin ningún ajuste las fórmulas de legibilidad del inglés a los textos escritos en castellano. Así lo hace, por ejemplo, la versión para esta lengua de Grammatik (Novell Inc., 1996) incluida en las herramientas lingüísticas (Proofing Tools) de Microsoft Office 97. A modo de caso práctico, consideremos el siguiente fragmento breve del cuento “La mariposa y la llama” de Azorín, tomado de la gramática de Alonso y Henríquez Ureña (1938):

Es el otoño. Las arboledas se tiñen de un amarillo pálido; luego el amarillo es más intenso; luego, el matiz es de oro viejo. Blanca ha salido de Madrid para hacer, en automóvil, el viaje a León. Desea ver, en esos días melancólicos del año, la plazoleta. El automóvil, poderoso, camina rápidamente. Blanca contempla, a lo lejos, la silueta de las montañas, y no piensa en nada. A la salida del Guadarrama, un accidente hace detenerse el coche. No ha pasado cosa mayor; los viajeros no han sufrido ningún daño; pero es preciso volver a Madrid para reparar los desperfectos del coche. ¿Podrán a la mañana siguiente reanudar el viaje los distinguidos viajeros? La jornada ha comenzado mal. Dos días después del retorno a Madrid, Blanca recibe un telegrama de París. Asuntos urgentes reclaman su presencia en la capital de Francia. El viaje a León queda aplazado indefinidamente; pero Blanca piensa en la vieja ciudad y con los ojos del espíritu ve la reducida plazoleta.

Hay que señalar que se trata de un fragmento de texto seleccionado por los profesores Alonso y Henríquez Ureña como ejercicio de una de las primeras lecciones de su curso de gramática castellana para la enseñanza secundaria, por lo que cabría esperar que su nivel de dificultad no fuese demasiado elevado. Sin embargo, la evaluación de Grammatik no puede ser más negativa, con un valor para IL de 27 equivalente a un grado muy difícil de lectura:

Cómputo			
Caracteres	773	Oraciones cortas	12
Sílabas	329	Oraciones largas	3
Palabras	165	Oraciones simples	12
Oraciones	13	Palabras largas	33
Párrafos	1		
Promedios			
Sílabas por palabra			1.95
Palabras por oración			12.69
Oraciones por párrafo			13.0
Legibilidad			
Grado en la escala Flesch (100 = muy fácil)			27
Voz pasiva (% de oraciones verbales en forma personal)			0
Complejidad oracional (100 = muy complejo)			3
Complejidad del vocabulario (100 = muy complejo)			59

Según este programa de verificación, el fragmento presentaría graves dificultades de comprensión lectora y, por tanto, tal vez no debería estar incluido en una antología de textos destinada a enseñar castellano a estudiantes de secundaria. Por si cabe alguna duda sobre los cómputos empleados, he aquí la explicación del índice de legibilidad de Flesh en los textos de ayuda de Grammatik:

Ayuda gramatical y estilística

Buscar | Temas de Ayuda | Anís | Opciones

◆ **Grado en la escala Flesch**

Este número es alto cuando el documento resulta fácil de leer y bajo cuando la redacción es complicada. Muchas organizaciones requieren que sus redactores o especialistas escriban a un cierto nivel en esta escala. Interprete el nivel de su documento según las siguientes especificaciones:

Nivel	Dificultad de lectura
90-100	Muy fácil
80-90	Fácil
70-80	Regular
60-70	Estándar
50-60	Un poco difícil
30-50	Difícil
0-30	Muy difícil

La fórmula de Flesch es:
 $206.835 - (1.015 \times \text{promedio de palabras/oraciones}) - (84.6 \times \text{promedio de sílabas/palabras})$

Por supuesto, el problema no radica en el texto, sino en la fórmula de legibilidad aplicada. Para paliar estos desajustes, distintos autores han tratado de adaptar al castellano la celeberrima fórmula de Flesh para el inglés. Así, Fernández Huerta (1959), argumentando que “en dicha lengua [inglesa] hay mayor disminución que en la española respecto del número de sílabas en las palabras de más uso”, sugiere sustituir por 0,6 el factor de ponderación de la longitud de palabra, en vez del 0,846 propuesto por Flesh para el inglés. De este modo, la fórmula de Flesh adaptada al castellano quedaría como sigue:

$$IL = 206,835 - 0,60 SP - 1,015 PO$$

El índice de legibilidad así obtenido se interpretaría mediante la misma tabla de equivalencias propuesta por Flesh para su fórmula. Así, según nuestros cálculos, aplicando la fórmula de Fernández Huerta al fragmento de Azorín, con $SP=203$ y $PO=9,167$ –valores ligeramente distintos a los detectados por Grammatik–, obtenemos un IL de 76 (redondeado a las unidades), es decir, un nivel de legibilidad más bien fácil.

Otro ejemplo de fórmula de legibilidad para el castellano es la presentada por Crawford (1989). Esta fórmula calcula el índice de legibilidad (IL) de un texto en castellano a partir de la media de sílabas por 100 palabras (SP) y la media de oraciones por 100 palabras (OP), según el cómputo:

$$IL = - 0,205 OP + 0,049 SP - 3,407$$

El valor de IL resultado de la aplicación de esta fórmula de Crawford representa los años de escolarización estimados para comprender sin dificultades el texto analizado. En el fragmento de Azorín (con $OP=10,909$), obtendríamos el valor de 4,305 para IL, lo que contribuiría a sancionar la selección del texto en un libro de enseñanza secundaria.

Finalizaremos este repaso a las fórmulas de legibilidad postuladas para el castellano con la presentada por Spaulding (1956). Esta fórmula calcula el índice de legibilidad (IL) de un texto a partir de la media de palabras por oración (PO) y del porcentaje de palabras presentes en el texto que no aparecen en una lista de 1.500 palabras de uso frecuente en la lengua (DT), según el siguiente cálculo:

$$IL = 1,609 PO + 331,8 DT + 22,0$$

Debe notarse que el valor de DT se obtiene dividiendo el número de palabras “difíciles” (o sea, ausentes de la lista) por el total de palabras del texto. Los resultados de la fórmula de Spaulding deben interpretarse mediante la siguiente tabla:

<i>IL</i>	<i>Grado de dificultad</i>
121 o más	Excepcionalmente difícil
101-120	Difícil
81-100	Moderadamente difícil
61-80	Fácil
40-60	Muy fácil
39 o menos	Primera enseñanza

Hemos aplicado esta fórmula al fragmento de Azorín empleando la lista de 5.000 palabras más frecuentes del castellano de Juilland y Chang-Rodríguez (1964), en lugar de la lista de 1.500 palabras sugerida por Spaulding. Esta sustitución se justifica por razones operativas específicas: mientras que la lista de Spaulding sólo contiene formas léxicas lematizadas no flexionadas (el infinitivo de los verbos, el masculino singular de los sustantivos y adjetivos, etc.), el diccionario de frecuencias de Juilland y Chang-Rodríguez contiene también las formas léxicas flexionadas, por lo que se puede consultar directamente. Hay que tener en cuenta, sin embargo, que esta modificación de la fórmula de Spaulding produce como resultado índices de legibilidad algo menores que la fórmula original, ya que las palabras de un texto poseen mayor probabilidad de estar incluidas en la lista de Juilland y Chang-Rodríguez, que de estarlo en la lista de Spaulding. Con esta adaptación, se obtiene un valor de 68,924 para *IL*, confirmando de nuevo la facilidad de lectura del texto analizado.

NOTAS

Este trabajo, basado parcialmente en la investigación recogida con mayor amplitud y profundidad en Gómez Guinovart (1999), ha sido realizado dentro del proyecto XUGA 20402B97 (Xunta de Galicia), y con el soporte académico y personal de los miembros del Área de Lingüística General y del Seminario de Lingüística Informática de la Universidade de Vigo.

REFERENCIAS BIBLIOGRÁFICAS

- Alonso, A. y P. Henríquez Ureña. 1938. *Gramática castellana*. Buenos Aires: Losada.
- Biber, D. 1988. *Variation across Speech and Writing*. Cambridge: Cambridge University Press.
- Biber, D. 1989. "A Typology of English Texts". *Linguistics* 27 (1): 3-43.
- Biber, D. 1995. *Dimensions of Register Variation: A Cross-Linguistic Comparison*. Cambridge: Cambridge University Press.

- Chomsky, N. 1965. *Aspects of the Theory of Syntax*. Cambridge: The MIT Press.
- Coseriu, E. 1981. "Los conceptos de 'dialecto', 'nivel' y 'estilo de lengua' y el sentido propio de la dialectología", *Lingüística Española Actual* 3 (1): 1-32.
- Crawford, A. 1989. "La comprensibilidad de textos en español del nivel primario: fórmula y gráfico". *Leer en la escuela: nuevas tendencias en la enseñanza de la lectura*. Ed. Fundación Germán Sánchez Ruipérez. Madrid: Pirámide. 274-283.
- Crystal, D. y D. Davy. 1969. *Investigating English Style*. Londres: Longman.
- El País. 1994. *Libro de estilo*. Madrid: Ediciones El País.
- Farr, J., J. Jenkins y D. Paterson. 1951. "Simplification of Flesh Reading Ease Formula". *Journal of Applied Psychology* 35: 333-337.
- Fernández Huerta, J. 1959. "Medidas sencillas de lecturabilidad". *Consigna* 214: 29-32.
- Flesh, R. 1948. "A New Readability Yardstick". *Journal of Applied Psychology* 32: 221-233.
- Gómez Guinovart, J. 1993. "Análisis lingüístico automatizado de un corpus de textos periodísticos en soporte informático". *Actas del IX Congreso de Lenguajes Naturales y Lenguajes Formales*. Ed. C. Martín Vide. Barcelona: PPU. 323-329.
- Gómez Guinovart, J. 1999. *La escritura asistida por ordenador: problemas de sintaxis y de estilo*. Vigo: Universidade de Vigo (Servicio de Publicacións).
- Gregory, M. 1967. "Aspects of Varieties Differentiation". *Journal of Linguistics* 3: 177-198.
- Guerra, R. 1983. "Estudio estadístico de la sílaba en español". *Estudios de fonética I*. Eds. M. Esgueva y M. Cantarero. Madrid: CSIC. 9-112.
- Gunning, R. 1952. *The Technique of Clear Writing*. Nueva York: McGraw-Hill.
- Halliday, M. A. K. y R. Hasan. 1976. *Cohesion in English*. Londres: Longman.
- Heidorn, G. E., K. Jensen, L. A. Miller, R. J. Byrd, y M. S. Chodorow. 1982. "The EPISTLE Text-Critiquing System". *IBM Systems Journal* 21 (3): 305-327.
- Hernanz, M. LI. y J. M. Brucart. 1987. *La sintaxis: Principios teóricos. La oración simple*. Barcelona: Crítica.
- Holmes, D. I. 1994. "Authorship Attribution". *Computers and the Humanities* 28: 87-106.
- Holmes, D. I. 1998. "The Evolution of Stylometry in Humanities Scholarship". *Literary and Linguistic Computing* 13 (3): 111-117.
- Jensen, K., G. E. Heidorn y S. Richardson, eds. 1993. *Natural Language Processing: The PLNW Approach*. Boston: Kluwer.
- Juilland, A. y E. Chang-Rodríguez. 1964. *Frequency Dictionary of Spanish Words*. Londres: Mouton.
- Kessler, B., G. Nunberg y H. Schütze. 1997. "Automatic Detection of Text Genre". *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*. 32-38.

- Klare, G. R. 1963. *The Measurement of Readability*. Ames: Iowa State University Press.
- Klare, G. R. y B. Buck. 1954. *Know your Reader: The Scientific Approach to Readability*. Nueva York: Hermitage.
- La Vanguardia. 1986. *Libro de redacción*. Barcelona: La Vanguardia-TISA.
- McGowan, S. 1992. "Ruskin to McRuskin: Degrees of Interaction". *Computers and Writing: State of the Art*. Eds. P. O'Brian y N. Williams. Boston: Kluwer. 297-318.
- McMenamin, G. R. 1993. *Forensic Stylistics*. Amsterdam: Elsevier.
- Núñez Ladévéze, L. 1979. *El lenguaje de los media: Introducción a una teoría de la actividad periodística*. Madrid: Pirámide.
- Patterson, W. y H. Urrutibéheity. 1975. *The Lexical Structure of Spanish*. La Haya: Mouton.
- Rodríguez Diéguez, J. L. 1989. "Predicción de la lecturabilidad de textos en castellano: una propuesta y sugerencias". *Leer en la escuela: nuevas tendencias en la enseñanza de la lectura*. Ed. Fundación Germán Sánchez Ruipérez. Madrid: Pirámide. 284-310.
- Rojo, G. 1991. "Frecuencia de fonemas en español actual". *Homenaje ó profesor Constantino García*. Eds. M. Brea y Fco. Fernández Rei. Santiago de Compostela: Universidade de Santiago de Compostela. 451-467.
- Spaulding, S. 1956. "A Spanish Readability Formula". *Modern Language Journal* 40 (8): 433-441.
- Vázquez-Ayora, G. 1977. *Introducción a la traductología*. Washington: Georgetown University Press.
- Widdowson, H. 1975. *Stylistics and the Teaching of Literature*. Harlow: Longman.