

HACIA UN ANÁLISIS GLOBAL DEL TEXTO: CARACTERÍSTICAS LÉXICAS, SINTÁCTICAS Y SEMÁNTICO-PRAGMÁTICAS*

Introducción

Los análisis de lengua escrita tienen ya amplia tradición bibliográfica, especialmente estimulante a partir del desarrollo de los estudios llamados de “análisis de texto” o “análisis del discurso”. Hacia la década de los setenta surgen figuras muy importantes en este tipo de investigación. Hunt (1965), van Dijk (1972), Halliday y Hasan (1976) son trabajos pioneros que han ido desarrollando parámetros específicos para la caracterización de los textos. Por otro lado, las publicaciones más recientes de la gramática funcional y las posteriores de estos mismos autores han ido proporcionando nuevos instrumentos para la descripción de la estructura textual (Hunt 1970, Beaugrande y Dressler 1981, van Dijk 1985, Halliday y Hasan 1989, Chafe 1994, Givón 1983, por citar algunos de los más representativos) Hoy a este tipo de trabajos se añaden los relacionados con los estudios de los grandes corpora. Estos requieren equipos electrónicos y metodologías muy especializadas y pueden describir los rasgos más idiosincrásicos de amplias parcelas de un texto o de la lengua. Con estos grandes corpora se elaboran hoy en día gramáticas de lengua oral¹ y diccionarios léxico-gramaticales.² Esta metodología se puede utilizar igualmente para

¹ Las gramáticas de lengua oral han cogido gran auge últimamente, su objetivo es principalmente didáctico. Pretende ofrecer al profesor los instrumentos adecuados para “cernir de manera óptima problemas precisos” (De Kock 1990:7) del español. Se apoyan en las concordancias recogidas de los grandes corpora escritos u orales. Con ellas se pueden analizar aspectos o construcciones particulares de una lengua, además de su léxico, que puede analizarse según aparece en índices y concordancias. El español cuenta con el proyecto de *Gramática española: enseñanza e investigación*, de Josse de Kock y equipo, que publica la Universidad de Salamanca. Hasta ahora han salido a la luz varios índices y concordancias de textos escritos y estudios particulares sobre formas verbales, pronombres, orden de los sintagmas, etc. (Véase De Kock 1990).

² Para un análisis de los actuales esfuerzos y problemas en estos trabajos véase Wilks *et alii* (1990). Los autores ofrecen una visión especializada sobre las técnicas de composición de corpora y las confecciones posteriores de diccionarios y lexicones electrónicos. Los corpora más conocidos en lengua inglesa son el LOB Corpus, el London-Lund Corpus de Spoken English y el de Brown University. Para una descripción detallada de ellos, véase Biber (1988: 65-71). En español contamos hoy día con grandes bases de datos informatizadas destinados al análisis automático de la lengua, por ejemplo, en la Universidad Autónoma de Barcelona, el BCL (Base de Conocimiento Léxico) y el BCS (Base de Conocimiento Sintáctico), que contiene un diccionario electrónico con más de 600,000 formas, permiten ya cierto grado de análisis automático de textos. Entre los conjuntos textuales más representativos se encuentran el archivo de textos hispánicos de la Universidad de Santiago (ARTHUS), con un total aproximado de 1,500,000 formas de textos procedentes de España y América pertenecientes a distintas etapas de la historia del español (la parte contemporánea constituye la Base de Datos Sintácticos del

estudios particulares de menor alcance: análisis de cualquier tipo de texto, autor o época, escritos de escolares, etc.³

Estas investigaciones se apoyan en las relaciones entre forma y significado tal como se han ido desarrollando en las últimas décadas. El léxico se ha descrito como una zona donde confluyen fenómenos lingüísticos de otros niveles, por ejemplo, sintácticos y pragmáticos. En este trabajo queremos hacer algunos planteamientos sobre ellas e insistir en la relativa comodidad con que hoy en día se pueden llevar a cabo los análisis estadísticos de grandes proporciones gracias a las nuevas técnicas que se han desarrollado en los programas de análisis de texto y recuento de datos.

Los acercamientos de análisis léxico más frecuentados en la actualidad permiten coordinar el léxico con la sintaxis. Los modelos más recientes relacionan léxico, sintaxis y discurso. Biber (1988), Sinclair (1991) y Stubb (1996) podrían considerarse muestras significativas en ese sentido, los tres presentan metodologías elaboradas que se apoyan en los patrones de asociación de palabras. En ellos el léxico se estudia como el punto de intersección de la sintaxis y la semántica. La palabra, además de permitir una interpretación por sí misma, se estudia con la perspectiva más amplia de verla en el conjunto de las formas con las que recurrentemente se asocia. Desde hace mucho tiempo se sabe que no es la palabra aislada la que transmite la información, sino la sucesivas combinaciones de ellas en los textos. Hoy tal vez estamos en posición de decir que existe la metodología capaz de probarlo.

La relación entre léxico y discurso no es nueva es una de las formas tradicionales de acercarse a los textos, lo nuevo es el acercamiento analítico aplicado y el equipo utilizado. El equipo electrónico ha permitido mayor alcance en los estudios. Este tipo de investigación tiene como primer objetivo la recogida de grandes conjuntos de texto y la finalidad inmediata del análisis automático de ellos es la confección de diccionarios de uso.⁴ Las técnicas analíticas

español actual, BDS); el ONE71 de la Universidad de Gotemburgo; y el ADMYTE de la Universidad Autónoma de Madrid, entre otros. La Real Academia de la Lengua Española cuenta con varios grandes corpora en marcha, entre los que se encuentran el CREA y otra gran base de datos para el análisis sociolingüístico del español actual hablado de España y América, específicamente el PRESEEA recoge las entrevistas puertorriqueñas.

³ La búsqueda de medidas cuantitativas y cualitativas que den cuenta de modo adecuado del crecimiento sintáctico que experimenten los niños en el curso de su desarrollo natural y su aprendizaje escolar ha sido y es objeto de preocupación constante entre los especialistas. En cuanto a las medidas cuantitativas, los índices de Hunt han probado ser los más provechosos en el análisis y la evaluación lingüística y han recibido el beneplácito de la gran mayoría de los educadores. Todos están de acuerdo, a su vez, de que se necesitan medidas complementarias para la descripción más adecuada. Hoy la disponibilidad del soporte informático acrecienta las posibilidades de descripciones exhaustivas de grandes conjuntos de textos.

⁴ Los primeros intentos de este tipo de análisis computarizado son Olney (1967) y Kucera y Francis (1967). Según Wilks *et alii* (1996:6), que cita al primero, estos trabajos se hacían con las viejas tarjetas de cómputos perforadas por duplicado para evitar errores.

permiten recoger cada forma en muchos contextos reales. Conforme se han ido perfeccionando los instrumentos operacionales, según las nuevas técnicas y programas de soporte informático se han hecho más flexibles y eficaces, se han ido extendiendo los objetivos de estudio, los diccionarios léxico-sintácticos son hoy investigaciones muy poderosas que permiten describir el comportamiento del léxico de una lengua en todas sus capacidades combinatorias. El examen sistemático de un cuerpo representativo de textos permite obtener una calidad de evidencia que no se había conseguido antes. Esta sistematicidad en el análisis hace posible conocer qué palabras son de mayor uso y cuáles ya han desaparecido de la lengua, incluso cuán apropiado puede ser el significado que se le asigna a una forma dada (Sinclair 1991: 37). Los esquemas de organización gramatical y semántica de cada entrada léxica, es decir su microestructura, se traza según los patrones reales de asociación que cada una de las unidades establece con las otras unidades de la lengua.

La posibilidad de obtener estos patrones de co-selección léxica ha abierto las puertas a análisis textuales de otro tipo. Con ellos se puede describir la particularidad que presenta un autor o una obra en específico, porque el análisis de las combinaciones de palabras permite identificar las relaciones establecidas entre ellas a distintos niveles descriptivos. Los esquemas repetitivos de determinadas asociaciones de palabras arrojarán mucha luz para su interpretación. Estos hechos cobran ahora especial importancia y son motivo de trabajos particulares que analizan las connotaciones culturales que aportan estas secuencias repetidas de palabras.⁵ Describir cómo asociaciones determinadas de formas, por ejemplo, un tiempo verbal con determinada clase de verbo, selecciones preposicionales específicas, usos preponderantes transitivos o intransitivos de un verbo, modalidades de las acciones, asociaciones de rasgos semánticos, etc., caracterizan el discurso de un texto, ha pasado a ser tema de gran interés en las recientes investigaciones de todas las lenguas. Interés justificado si se tiene en cuenta que el análisis no proviene ya, como los anteriores, de la motivación particular que pueda tener el investigador por una determinada combinación de palabras o formas, aunque este tipo de trabajo no está excluido, sino que parte de un acercamiento metodológico que analiza simultáneamente todo el texto o textos y garantiza descubrir las combinaciones más significativas en él. Por otro lado, parece haber consenso general respecto a que cuando se estudian las obras bajo perspectivas sintácticas o léxicas por separado, que es como nos hemos acostumbrado a hacerlo, las relaciones entre ambos no siempre salen a la luz y gran parte de la idiosincrasia de ese discurso o corpus puede haber quedado sin recoger. Los acercamientos actuales permiten describir con cierta seguridad la naturaleza de los textos y la mayor o menor innovación que

⁵ Véase, por ejemplo, en español la oposición entre las dos denominaciones 'España' y 'la nación española' (García Santos 1987).

representan frente a la lengua en general o frente a otro conjunto mayor de textos al que pertenezca.

Así pues, en general, las listas léxicas que se construyen de este modo tanto para diccionarios, como para análisis interpretativos textuales unen forma con significado apoyados por evidencia real exhaustiva y esta circunstancia asegura su legitimidad. Las descripciones de las unidades son, por un lado, muy concretas porque se apoyan en la evidencia real encontrada en el texto y, por otro, mucho más abstractas porque se obtienen de todas las menciones que presentan los textos. La caracterización de cada unidad requiere objetivar y abstraer el patrón general compartido por todos los casos, con ello se garantiza su sistematicidad.

Este tipo de trabajo tiene amplia tradición en la lingüística británica y tiene como base "la teoría contextual del significado" de Firth 1935, 1957.⁶ Aquí les queremos ofrecer un panorama general de las posibilidades teóricas y metodológicas que ofrece la lingüística textual actual. Todos los análisis están orientados a la identificación de los índices o señales que, de algún modo, revelan la idiosincrasia del texto o textos que se analizan. El objetivo de análisis en todos ellos es recoger sus peculiaridades, sean de tipo léxico, sintáctico o funcional; y trátense de textos particulares, de conjunto de ellos o de los representativos de una lengua.

Metodología

Para los análisis de grandes conjuntos de texto existen programas de soporte informático muy refinados. Además de utilizar las medidas estadísticas basadas en la relación entre palabras diferentes y el total en el texto, la proporción entre palabras plenas y vacías, etc. que dan la apreciación general de la estructura del léxico, tienen, como ayudas indispensables, programas de *concordancias*. Estos programas buscan las unidades del texto y las presentan en el centro de una página con las que la acompañan en la línea a cada lado. Proporcionan lo que se ha llamado palabras en contexto. Algunos programas las pueden presentar en orden alfabético según la palabra que le sigue a la derecha, e incluso puede contar las variantes, lo cual permite recoger las de mayor representatividad en el texto. Son programas de kwic ("key word in context") muy mejorados. Los programas de análisis de texto que incluyen concordancia abundan, entre ellos se encuentran Longman Mini-Concordancer, escrito para MS-DOS por Brian Chandler; CONC del Academic Computing de Dallas,⁷

⁶ La teoría contextual del significado de Firth (1935, 1957) establece que el lenguaje debe ser estudiado en ejemplos auténticos de uso, y éstos en textos completos, que, a su vez, se analizarán comparativamente en todo un cuerpo de textos. Para Firth el significado no es sólo un fenómeno mental sino que tiene que ser analizado con datos objetivos de los textos. La lingüística debe analizar por lo tanto la forma y el significado, que son inseparables y están entrelazados en los textos (Stubbs 1996: 23)

⁷ Este programa está diseñado para facilitar el estudio intensivo de un texto produciendo la lista de

BUSCA de Lee Hartman,⁸ EXÉGESIS de Raúl Avila,⁹ WordSmith de Mike Scott¹⁰ y otros más. Muchos de ellos se pueden conseguir ahora por Internet y algunos son libres de costo (Para más información, cf. Morales 1989).

Análisis del léxico

Las funciones más básicas de este tipo de programa son los recuentos de palabras. Con ellos se obtiene la estructura léxica del texto: índices de densidad y riqueza léxica, concentración del vocabulario, distribución, etc. Dependiendo del programa que se utiliza y de los propósitos del estudio, los textos tendrán que estar o no previamente codificados. Con este recuento se hace el listado básico ordenado por frecuencia, lo cual permite ya comparar el léxico del texto que se analiza con cualquier otro publicado que sea representativo de esa lengua, por ejemplo, los léxicos básicos o disponibles o cualquier otro conjunto de palabras de una lengua. Se obtienen igualmente otras medidas como la proporción entre el total de palabras y el número de palabras diferentes, la gradación de estas mismas en el texto, la representatividad numérica de las palabras de frecuencia 1 (hapax legomena), etc.; igualmente el número de palabras por oración, datos que permiten, como primer acercamiento, la obtención de unos cuantos índices asociados a la riqueza del léxico y a su complejidad. Junto a ellos se puede obtener una primera caracterización gramatical si las listas aparecen enriquecidas con clases de palabras. Para ello el texto obligatoriamente debe haber sido codificado con anterioridad.¹¹

palabras que ocurren en él con una sección de contexto precedente y siguiente. Produce, también, la lista de palabras diferentes del texto indicando todos los lugares en que aparece. Tanto la concordancia como el índice pueden hacerse de palabras seleccionadas previamente o de los totales; con texto en secuencia o con texto interlineal. Produce, además, concordancias de morfemas.

- ⁸ Este programa está basado en DOS, busca patrones específicos de palabras o cadenas de letras en los textos. Analiza oración por oración apoyándose en las marcas de puntuación. Las palabras pueden definirse de acuerdo con sus letras iniciales, finales o cualquier grupo interior. Se ha usado ya con el corpus español de los textos argentinos y chilenos de CRLEC (Corpus de Referencia de la Lengua Española Contemporánea). Está ahora disponible para análisis del español e inglés.
- ⁹ Este programa pertenece al proyecto de estudio de los medios de comunicación que dirige desde México Raúl Avila. Cuenta con cálculos estadísticos que ofrecen los índices de riqueza léxica y similares y un programa de concordancias ("Contexto"). Es intercomunicativo y su capacidad de análisis va aumentando según los distintos usuarios lo utilizan e incluyen en él la información léxica necesaria para sus descripciones.
- ¹⁰ Los autores definen este programa como el "Swiss Army knife of lexical analysis", un conjunto integrado de programas que analizan cómo se comportan las palabras en el texto. Esta hecho especialmente para lingüistas y profesores de lengua. Tiene tres instrumentos principales: lista de palabras, concordancias y palabras clave. Con el primero produce listas de palabras (tanto de palabras diferentes, como "running words") e índices estadísticos basados en la relación entre ellas. Cuenta las oraciones y su longitud. Con el segundo crea concordancias, encuentra las de determinadas palabras, identifica los contextos iguales y los más frecuentes (colocaciones) y produce gráficas relacionadas con esta información. Las palabras clave identifican a las más frecuentes y sus asociaciones más representativas. Ofrece también gráficas de ellas.
- ¹¹ La codificación de los textos depende de los propósitos del estudio y del tipo de programa que se está

Este primer acercamiento cuantitativo tiene el inconveniente de que los índices no son muy significativos aisladamente. Necesitan la comparación con los obtenidos en textos diferentes o en la lengua general. La riqueza léxica y las medidas asociadas a ella son índices relativos, no dicen nada por sí solos. Una buena aproximación se obtiene si distinguimos previamente las distintas partes (capítulos, contenidos, etc.) que constituyen el texto que queremos analizar. De ese modo se pueden establecer comparaciones entre los índices de cada una de ellas y, a la vez, con los índices generales del conjunto. Esta operación ofrecería la estructura léxica del texto en cuanto a su constitución interna. También podrían describirse períodos diferentes de un mismo texto o el discurso de autores representativos de un período o de una tendencia.

Este análisis léxico se enriquece si se tienen en cuenta las llamadas "palabras clave" (key words).¹² Las palabras clave son las palabras que por sus rasgos de frecuencia difieren del conjunto, bien por tener frecuencia superior al resto. Es una operación que llevan a cabo la mayoría de los programas citados, las identifican por contraste y algunos de ellos muestran gráficamente donde ocurren estas palabras en el texto. Desde luego se pueden listar aparte e, incluso, pueden aparecer con las palabras que en su sección de texto están asociadas por frecuencia. Dato que podría indicar que existe cierta relación semántica. Las palabras clave son, sin lugar a dudas, palabras relacionadas con el contenido semántico que ofrece el texto y con la ideología que representa; pero un estudio cuidadoso de ellas puede mostrar no sólo los aspectos significativos relacionados con el tema tratado, sino índices de otro tipo de significado que trasciende el ofrecido en el texto o textos.¹³ Indudablemente las palabras clave recogerán además los rasgos de estilo del autor, época o género.

Otro modo de enriquecer el análisis a este nivel es contrastar los datos de determinadas palabras clave cuando ellas encierran la posible competencia entre dos lenguas. Los textos que, como los puertorriqueños, por ejemplo, pueden presentar anglicismos, tienen aquí la posibilidad de un acercamiento muy provechoso, pues se puede ver el comportamiento y uso de la palabra

utilizando. El EXÉGESIS, por ejemplo, exige una codificación muy minuciosa. Normalmente cuando se trata de un análisis gramatical conviene tener claramente identificadas las clases de palabras y algunas lexías. Siempre es conveniente deshacer las ambigüedades semánticas. (Para unas guías sobre la codificación del español, véase Morales 1987.) Programas como el WordSmith permiten ser más flexibles con la codificación, ya que realizan muchas operaciones de estadística básica general y posteriormente se pueden identificar las palabras por el contexto de las concordancias en los análisis más específicos.

¹² Palabra clave es "todo término de frecuencia y distribución suficientes cuyo papel es cristalizar un cierto número de ejes semánticos mayores del discurso.." (Maingueneau 1982:9).

¹³ Las palabras clave dirigen al investigador al contenido doctrinal o ideológico de un texto y marcan la posición del autor o autores respecto a esos hechos u otros más generales de tipo social o cultural. El análisis computarizado de las mismas se establece comparando dos listas creadas previamente, una de ellas actúa como lista de referencia y la otra es la que se analiza. Generalmente la primera se hace con la totalidad de los materiales o con materiales modelo.

patrimonial en oposición al préstamo ofreciendo al programa una lista previa de las palabras que se quieren analizar.

Análisis del discurso

Primera fase

COLOCACIONES

El primer acercamiento al análisis cualitativo del discurso nos lo ofrecen estos mismos programas con las "concordancias" o palabras en contexto. Estos programas se han ido enriqueciendo en los últimos años y ahora permiten obtener la palabra con los contextos que se le asocian con mayor frecuencia. Se trata de obtener las "colocaciones" o conjuntos de "palabras en su habitual compañía" (Stubbs 1996:173). Identifican, así pues, en las concordancias, las frases o grupos de palabras más comunes y, con ello, las que pudieran ser más representativas como patrones léxico-sintácticos de la lengua y, a su vez, más significativas en un acercamiento semántico-ideológico.

Las colocaciones se fundamentan en una doble teoría de interpretación léxica que postula que en un texto las palabras se suceden *at random* con elección libre u ocurren agrupadas en "idioms" (Sinclair 1991).¹⁴ En este último caso el texto se construye por la elección de bloques de palabras pre-construidos. Las colocaciones reflejan el principio del *idiom* porque parte de que en ocasiones las palabras se eligen por parejas o por grupos mayores, no aisladas, una a una. El hecho es que aunque no podamos decir, ni mucho menos, que un texto está constituido únicamente de lexías o frases hechas, lo cierto es que el margen de libertad en la asociación de las palabras es menos del que a primera vista cabría pensar. Determinadas palabras co-ocurren con frecuencia significativa en él, aunque eso sí, muestran niveles diferentes y graduales de asociación. Se pueden distinguir diferentes clases de colocaciones de una forma dependiendo del nivel de frecuencia de las palabras con las que se asocia.¹⁵

Estas combinaciones permiten obtener tanto interpretaciones formales, como semántico-pragmáticas. Las primeras recogen cada palabra con su entorno léxico-gramatical y proporcionan las microestructuras de las entradas de los diccionarios de uso de los grandes corpora. Basta pensar en el problema de los regímenes preposicionales y verbales. Estos según las asociaciones más frecuentes entre él y el complemento, o la preposición, etc., pueden venir ahora categorizados de acuerdo con sus características reales de uso. El diccionario Collins COBUILD (Sinclair et alii 1987) se apoya en un corpus de veinte millones de

¹⁴ Queremos mantener la palabra "idiom" del autor porque es un término general que comprende tanto construcciones preposicionales o conjuntivas casi fijas, como otras asociaciones de mayor flexibilidad.

¹⁵ Sinclair (1991) define la colocación como la manera en que se "colocan" las palabras con respecto a otras. Señala que cuando dos palabras de diferente frecuencia se unen, la colocación tiene diferente valor para cada una de las dos palabras, es más significativa para la palabra menos frecuente.

palabras procedentes de libros, revistas, periódicos, conversaciones, etc. Para cada una de las entradas se han analizado muchas concordancias y se han tenido en cuenta las colocaciones que forman. Según los autores se ha desarrollado una teoría de reflexividad de la lengua que les permite recoger muchos ejemplos de la falta de adecuación de algunas de las microestructuras de los diccionarios tradicionales.¹⁶

También aquí la concordancia podrá sacar a flote muchos usos anglicados. Se podrán detectar las construcciones que constituyen lexías o colocaciones ajenas al español (anglicismos semánticos o sintácticos) simplemente recogiendo las concordancias de determinadas palabras sensibles a esos procesos. El análisis se basará en la frecuencia con que la palabra bajo estudio se asocia con distintos entornos semánticos o sintácticos. La ventaja del acercamiento estadístico es que un estudio de este tipo permite ser exhaustivos en la recogida de datos.

ANÁLISIS FACTORIAL

Resultados similares se pueden obtener con metodología un tanto diferente. Nos referimos a la de los análisis estadísticos factoriales (Biber 1988).¹⁷ Esta metodología permite, como la anterior, análisis cuantitativos y cualitativos de variación textual. Para ello se utiliza el instrumento de análisis factorial con que cuenta cualquier programa estadístico (SAS, SPSS, SPSS-X, etc.). Los análisis factoriales sirven para localizar el conjunto de rasgos que co-varían en un texto previamente codificado. Ya no se trata tanto de asociación de palabras, sino de los rasgos lingüísticos que se han seleccionado en el estudio. Se recogen las asociaciones entre determinados rasgos que co-ocurren o la complementariedad de ellos, si no van a aparecer asociados. En este caso está marcado negativamente en su correlación con el otro con el que se contrasta. El análisis factorial es multidimensional todos los rasgos se contrastan con los otros.

Los rasgos lingüísticos seleccionados pueden variar de estudio a estudio y dependen de las intuiciones o experiencia previa del investigador. Los más socorridos en todas las lenguas son tiempo y aspecto verbal, persona gramatical, nominalizaciones o formas asertivas, clases de verbos, voz pasiva, tipos de adverbios, tipos de adjetivos y clases léxicas.¹⁸

¹⁶ Con los resultados obtenidos en los análisis se mejoran las microestructuras de las entradas de los diccionarios. En muchos casos de palabras del inglés Sinclair *et alii* (1990) observa cómo los significados dados como primeras acepciones no se corresponden con la realidad del uso e, incluso, la subclase asignada a algunas entradas verbales debería cambiar. Estas circunstancias le permiten asegurar que los diccionarios tradicionales son "museum of rare and unusual language events" (Sinclair 1991:38)

¹⁷ "Factor analysis" es un instrumento estadístico de aproximación multidimensional a la variación textual. Ofrece las combinaciones de variables que se obtienen de la correlación matriz de todas ellas (Biber 1988: 79-80). Son los aspectos o rasgos lingüísticos del texto que presentan variación compartida.

¹⁸ Con los análisis factoriales se precisan las asociaciones de determinados rasgos o la complementariedad

Segunda fase

Con las colocaciones podemos obtener análisis un tanto más profundos ya propiamente cualitativos orientados a establecer las relaciones de sentido que se dan entre unidades determinadas. Así como en sus aspectos más formales reflejan los patrones de organización léxico-gramatical, en los pragmático-significativos, son manifestaciones de los esquemas ideológicos del texto. Debido a que las concordancias permiten recoger las asociaciones de mayor frecuencia, ello proporciona, por lo pronto, las combinaciones de palabras de mayor significatividad en los textos. La frecuencia que obtengan las distintas colocaciones serán índices importantes para medir la trascendencia cultural y social de los conceptos que concretizan. Igualmente sacan a relucir, la “modalidad” semántica de la asociación, es decir, el sentido que trasciende cada combinación particular cuando se observa el matiz significativo que acompaña a todas y que sobrepasa a cada una en particular. Una determinada palabra puede estar relacionada en mayor medida con circunstancias placenteras o con tristes, con hechos de una naturaleza o de otra, etc. (Sirvan de ejemplo Frawley 1987, Atkinson 1992).¹⁹

El investigador tendrá que discernir cuáles de estas asociaciones repetidas responden meramente a rasgos de estilo, no por ello menos importantes para determinar el texto, y cuando reflejan ejes ideológicos o semánticos. Estas posibilidades interpretativas resultan mucho más objetivas que los meros comentarios que se puedan ofrecer basados en la pericia analítica subjetiva de cualquier investigador. Ya son muy conocidas las críticas hechas a determinados trabajos en el sentido de que carecen de análisis básicos esenciales y de que sólo con unas cuantas construcciones seleccionadas subjetivamente se construye el andamiaje interpretativo (Sobre ello, Phillips 1989 y Bell 1991).

Por otro lado, desde perspectiva un tanto diferente, tradicionalmente el investigador ha estado buscando los índices que marcan la formalidad lingüística. Se postula que la lengua escrita es discurso planeado²⁰ y que, como tal, tiene unas características especiales. Se le atribuye estructura oracional más compleja con recursos lingüísticos de mayor formalidad; cierta autonomía respecto al contexto o situación de habla; autosuficiencia temática o explicitud;²¹

que puedan presentar (donde aparece uno no aparece el otro). La complementaridad está marcada negativamente. Estos índices de asociación permiten obtener las correlaciones que se dan entre los rasgos manejados.

¹⁹ Los autores llaman a este fenómeno “prosodia semántica” (semantic prosody) que se define en Stubbs (1996:173) como el rasgo semántico que se extiende sobre más de una unidad. Sinclair señala el contexto poco feliz (“unpleasant”) que siempre acompaña a *happen* (1991: 112).

²⁰ Las cualidades que lo definen son que ha sido pensado antes de escribirse y que su función principal es la referencial (Silversten 1976). Según Ochs (1979) el dominio lingüístico de un escritor no sólo se refleja en la mayor complejidad estructural y significativa de sus textos, sino en su capacidad para sustraerse del contexto inmediato.

²¹ Se dice que un texto tiene suficiencia temática o que es suficientemente explícito cuando su comprensión

y coherencia textual que le permita mantener la secuencia lógica de los eventos y referencias con la utilización adecuada de los recursos gramaticales.²²

Los análisis más provechosos sobre la complejidad oracional provienen de Hunt (1970), sus índices dan razón de esa complejidad. Sin necesidad de tener que realizar un trabajo tan pormenorizado y lento, muchos programas estadísticos, por ejemplo, el WordSmith, dan también índices del largo de la oración. Con ellos se pueden establecer los primeros perfiles de complejidad sintáctica. Estos perfiles se podrían enriquecer con análisis cualitativos que permitieran evaluar las marcas del discurso formal. Para ello se requieren análisis más específicos enfocados a unidades determinadas, por ejemplo, además de todo lo dicho con anterioridad, se podría considerar el nivel de formalidad que representa un texto. Los recursos que tradicionalmente se han asociado al discurso formal: gerundio en función adverbial; participio con valor explicativo (especialmente en cláusulas de participio); voz pasiva; determinados adverbios; y adjetivos atributivos. Son todos rasgos que se pueden recoger de las concordancias o análisis factoriales orientando la investigación hacia ellos y codificando el texto de acuerdo con los intereses del análisis. La importancia de estos rasgos radica en que son marcas con características comunicativas especiales, todas ellas son portadoras de información nueva, son modificadores libres. Es bien conocido que en las unidades que son elementos opcionales en la oración descansa el mayor peso de la información nueva que se transmite en un texto (Siewierska 1991).²³

Por otro lado, un texto dado puede estar más o menos ligado a su entorno, es decir puede presentar información que para su comprensión el lector necesite tener, en mayor o menor grado, cierto conocimiento previo de lo que se narra o describe en él. Estas características tienen que ver con la autonomía del texto y la autosuficiencia temática, cualidades que han recibido diferentes valoraciones, aunque en general se supone que el texto escrito las presentará en mayor medida. Son características que ayudan a establecer el sistema de relaciones semánticas del texto y su conexión con el mundo real exterior. Respecto a ello resulta provechoso analizar lo que se ha llamado "grupos nominales saturados", que son los nombres que aparecen en los textos y cuyas características formales no se explican desde un punto de vista estrictamente lingüístico

no descansa en la experiencia compartida entre hablante y oyente (o entre autor y lector) y es capaz de independizarse del contexto que le rodea.

²² En realidad se trata de análisis de cohesión que describen los rasgos superficiales de las relaciones de coherencia, porque éstas son demasiado abstractas y complejas y trascienden los acercamientos empíricos (Hasan 1984, Halliday y Hasan 1989). No todos están de acuerdo en que haya relación directa entre coherencia y cohesión y hay investigaciones cuyo propósito ha sido medir esa relación. En general la teoría se mantiene y ya son muchas las interpretaciones textuales que se han hecho siguiendo estas directrices.

²³ Estos planteamientos han permitido considerar que las oraciones de relativo no restrictivas y los participios y adjetivos explicativos son cualitativamente más importantes que los restrictivos en cuanto a su aportación a un texto determinado.

(Basire 1982).²⁴ Gozan de la condición del nombre propio y pueden aparecer solos y con un artículo definido. Este tipo de nominal recuerda a las entidades que en la escala de Prince (1981) reciben el nombre de "no usadas" (unused). La particularidad de estos nombres es que pertenecen al trasfondo cultural o al conocimiento compartido de los hablantes y se pueden presentar por primera vez sin necesidad de recursos especiales de presentación.²⁵ Así, surgen en el texto con artículo definido aunque no hayan aparecido previamente. La importancia de estos nombres en los análisis de texto viene dada porque su estado referencial depende del universo discursivo en que está inmerso. Nombrar como experiencia cultural compartida determinada entidad puede ser muy significativo en un texto particular y en un contenido específico. Desde luego ayudará a configurar los ejes semánticos que lo sostienen.

Otras categorías gramaticales que han recibido atención especial son las nominalizaciones. No tiene la misma estructura argumental: "invadieron la ciudad" que "la invasión de la ciudad". Aunque los significados pueden darse como equivalentes hay una diferencia de condición asertiva. La nominalización presenta los hechos de un modo implícito y la aserción verbal de la oración los asevera (Serriot 1986). El uso mayoritario de una u otra forma podría responder a factores estilísticos o a intención comunicativa. En cada caso la elección sería significativa y en conjunto podría dar un matiz especial a todo el texto.

Por otro lado la lingüística del texto está interesada en la identificación de los "índices" que han sido señalados por la teoría de adquisición del lenguaje como marcas de madurez lingüística. Los análisis de textos escolares tienen que basarse en ellos. Se postula que determinadas oraciones de relativo, específicamente las que funcionan como frases preposicionales; además de las no restrictivas, son de adquisición tardía.²⁶

En cualquier caso, sea el estudio de mayor o menor alcance, si se va a utilizar soporte informático, es indispensable hacer, como primer paso, una lista de frecuencia que dará la guía de las palabras importantes y a las que nos debemos acercar con más interés. Igualmente podría ser necesario hacer las concordancias de algunas de ellas y si es posible las frecuencias de las

²⁴ Dice el autor: "Ces termes 'auto-définis' constituaient en quelque sorte un ensemble de 'valeurs' sur lesquelles se serait établi un consensus. Or ce consensus est un mythe (même si le locuteur y croit lui-même, ce qui est vrai dans la majorité des cas), une illusion qu'il faut dévoiler" (Basire 1982:60).

²⁵ Las entidades que representan estos sintagmas nominales tienen rasgos muy especiales: por un lado, son identificables en el discurso, ya que forman parte de la experiencia compartida; y por otro, son nuevos en él, porque no han aparecido previamente en la secuencia textual.

²⁶ Esto sucede porque en estos casos, la preposición, que obligatoriamente acompaña a la oración, introduce la nueva dificultad que se deriva de la posición independiente que adquiere la oración subordinada cuando la modificación se efectúa a través de una preposición, lo que no ocurre en el caso de la subordinada que va simplemente yuxtapuesta. La preposición expresa la existencia de una relación y proporciona, a la vez, según la preposición que intervenga, información específica sobre el carácter de la relación. Esta información tiene que estar combinada con la que transmiten los dos elementos relacionados para que se entienda el significado de la construcción.

concordancias para completar el cuadro de posibilidades. Esto permite tener una base segura respecto a lo que conviene hacer y evita perderse en la inmensa cantidad de datos que pueden resultar en las concordancias.

Amparo Morales
Universidad de Puerto Rico
Recinto de Río Piedras

Bibliografía

- Atkinson, D. "The evolution of medical research writing from 1735 to 1985." *Applied Linguistics* 13,4. 1992: 337-74.
- Basire, B. "Groupes nominaux saturés et discours." *DRLAV* 27. (1982): 47-61.
- Beaugrande, R. y W.U. Dressler. 1981. *Introduction to text linguistics*. London/New York: Longman, 1982.
- Bell, A. *The language of News Media*. Oxford: Blackwell, 1991.
- Biber, Douglas. *Variation across speech and writing*. Cambridge: Cambridge Univ. Press, 1988.
- Chafe, Wallace. *Discourse, consciousness, and time*. Chicago: The University of Chicago Press, 1994.
- De Kock, Josse. *Gramática española. Enseñanza e investigación. Apuntes metodológicos*. Salamanca: Ed. Universidad de Salamanca, 1990.
- Dijk, T.A. van. *Some aspects of text grammar*. La Haya: Mouton, 1972.
- _____. ed. *Handbook of discourse analysis*, vol 1 y 2. Nueva York: Academic Press, 1985.
- Fernández Lagunilla, Marina. "Léxico y discurso sociopolítico español." En V. Demonte y B. Garza Cuarón (eds.). *Estudios de Lingüística de España y México*, 351-69. México: UNAM y Colegio de México, 1990.
- Firth, J.R. "The technique of semantics." *Transactions of the Philological Society*, 1935: 36-72.
- _____. *Papers in linguistics*. London: Oxford University Press, 1957.
- Frawley, W. Review article: T.A. van Dijk, ed. "Handbook of discourse analysis." *Language* 63,2. (1987): 361-97.
- García Santos, F. 1987. "El lenguaje político: en la segunda República y en la democracia." En M. Alvar (coord.) *El lenguaje político*, Madrid: Fundación Friedrich Ebert, 1987.
- Givón, Talmy, ed. *Topic continuity in discourse*. Amsterdam/Philadelphia: John Benjamins, 1983.

- Halliday, M.A.K. y R. Hasan. *Cohesion in English*. London/New York: Longman, 1976.
- . *Language, context and text: Aspects of language in a social-semiotic perspective*. Nueva York: Oxford University Press, 1989.
- Hunt, K.W. *Grammatical structure written at three grade levels*. NCTE Research Report, N°3. Urbana, Ill., 1965.
- . "Recent measures in syntactic development." En M. Lester (ed.) *Reading in applied transformational grammar*, Nueva York: Holt, Rinehart and Winston, 1970.
- Kucera, H. y W.N. Francis. *Computational analysis of present-day American English*. Providence: Rhode Island, 1970.
- Maingueneau, D. *Nouvelles tendances en analyse du discours*. Paris: Hachette Université, 1987.
- Morales, Amparo. "Diccionarios de frecuencia y segmentación en español." *Revista Española de Lingüística* 10,2. (1980): 467-74.
- . "Programas de concordancias para análisis de texto." *Lingüística* 1. 1989: 320-29.
- Olney, J. *Toward the development of computational aids for obtaining a formal semantic description of English*. SDC Technical Report Series (SP- 2766). Santa Mónica, CA: System Development Corp., 1967.
- Phillip, M. "Lexical structure of text." *Discourse Analysis Monograph 12*. Birmingham: English Language Research, 1989.
- Prince, Ellen. "Toward a taxonomy of given/new information." En E: Cole (ed.) *Radical Pragmatics*, Nueva York: Academic Press, 1981.
- Serriot, P. "Langue russe et discours politique soviétique: analyse des nominalisations." *Langages* 81. (1986): 11-41.
- Siewierska, Anna. "Functional Grammar." London/New York: Routledge, 1991.
- Sinclair, J. McH., ed. *Collins Cobuild English Language Dictionary*. (2ª ed, 1995) London: HarperCollins, 1987.
- Sinclair, J. McH., et al. *Collins COBUILD English Grammar*. London: Collins, 1990.
- . *Corpus, Concordance, Collocation*. Oxford: Oxford University Press, 1991.
- Stubbs, Michael. "The state of English in the English state: Reflections on the Cox Report." *Language and Education*, 3,4. (1989): 235-50.
- . "Collocations and semantic profiles: On the cause of the trouble with quantitative methods." *Functions of Language*, 2,1. (1995): 1-33.
- . *Text and corpus analysis*. Cambridge: Blackwell, 1996.
- Wilks, Yorick A., et al. *Electric Words*. Cambridge, MA: MIT Press, 1996.