2 Fundamentos lingüísticos

En este capítulo se describen las consideraciones lingüísticas que sirven de base al estudio del lenguaje y de la anáfora indirecta. Primero se describen brevemente los niveles de estudio del lenguaje, para ubicar el estudio de la anáfora indirecta entre ellos, y el concepto general del contexto. En segundo lugar se presenta el texto y sus propiedades: la adecuación, la cohesión y la coherencia; estas propiedades deben encontrarse en cualquier texto, y se identifica su interrelación y los mecanismos en que se apoyan.

2.1 Los niveles del lenguaje

El lenguaje para su estudio, análisis y procesamiento ha sido dividido en niveles de acuerdo a la cantidad y estructura (u organización) de los datos. Tomando en cuenta que el origen de la representación textual es la representación del habla, tenemos:

1. Fonética – estudia los sonidos del habla desde el punto de vista de sus características físicas o articulatorias que influyen en la generación de voz con diferente volumen y tonos.
2. Fonología – estudia el conjunto de relaciones de los sonidos en la generación e interpretación del habla. Se considera la rama de la lingüística que estudia los elementos fónicos, atendiendo a su valor distintivo y funcional [DRAE].
3. Morfología – estudia la estructura y relaciones de las agrupaciones de símbolos para la formación de palabras. Se considera la parte de la gramática que se ocupa de la estructura de las palabras [DRAE].
4. Sintaxis – estudia la estructura y relaciones en las agrupaciones de palabras para la formación de frases y oraciones. Se considera la parte de la gramática que enseña, por medio de un conjunto de reglas, a coordinar y unir las palabras para formar las oraciones y expresar conceptos [DRAE].

5. Semántica – estudia la interpretación del significado de las palabras tomando en cuenta el uso general y sus relaciones sintácticas en forma independiente del contexto del discurso [SIL].

6. Pragmática – estudia la interpretación y el significado de las palabras tomando en cuenta el contexto del discurso en el que se utilizan, incluyendo la intención supuesta del emisor y la del receptor [SIL]. En otras palabras, es el estudio de como las expresiones analizadas gramaticalmente interactúan en relación con el contexto de interpretación.

La interpretación de la correferencia y de la anáfora se apoya en los niveles 4 al 6 (sintáctico, semántico y pragmático). La interpretación de la anáfora indirecta en particular requiere apoyarse principalmente en el nivel 6 (pragmático).

2.2 El contexto del discurso

En general, se entiende por contexto del discurso *el conjunto de conocimientos y creencias compartidos por los interlocutores de un intercambio verbal y que son necesarios para* ***producir e interpretar*** *sus enunciados*.

En el estudio del contexto del discurso se reconocen tres componentes: el sociocultural, el situacional y el lingüístico.

El contexto sociocultural es la configuración de datos que proceden de condicionamientos sociales y culturales sobre el comportamiento verbal y su adecuación a diferentes circunstancias. Hay regulaciones sociales, por ejemplo, sobre cómo saludar o sobre qué tratamiento o registro lingüístico usar en cada tipo de situación.

El contexto situacional, es el conjunto de datos accesibles a los participantes de una conversación, que se encuentran en el entorno físico inmediato. Por ejemplo: para que el enunciado "*cierre la puerta, por favor*" tenga sentido, es necesario que haya ciertos requisitos o *presuposiciones que son parte de la situación* de habla: que haya una puerta en el lugar donde ocurre el diálogo, y que esté abierta.

El contexto lingüístico está formado por *el material lingüístico que precede y sigue a un enunciado*. En las actividades de lectura el contexto lingüístico es de gran importancia para

inferir palabras o enunciados que no conocemos. Vale la pena puntualizar aquí una diferencia entre el texto oral y el escrito; en el texto escrito el contexto lingüístico *incluye paulatinamente* la información necesaria para construir el contexto situacional en su proceso de generación o interpretación; esto se debe a que, entre emisor y receptor, no es posible: la interacción de los sentidos con elementos del entorno común; ni la posibilidad de solicitar una corrección o una aclaración en el proceso de comunicación, como ocurre en una conversación oral.

El contexto del discurso, necesario para el proceso de la ***producción e interpretación*** del lenguaje, puede imaginarse como un témpano flotante (iceberg) donde lo único perceptible o superficial considerado como **explícito** es: el conjunto de símbolos agrupados para formar unidades léxicas; estas unidades léxicas agrupadas en estructuras formando frases y oraciones; y el agrupamiento de oraciones para formar párrafos, que dan forma y estructura a un documento de texto; como se muestra en la figura 1.

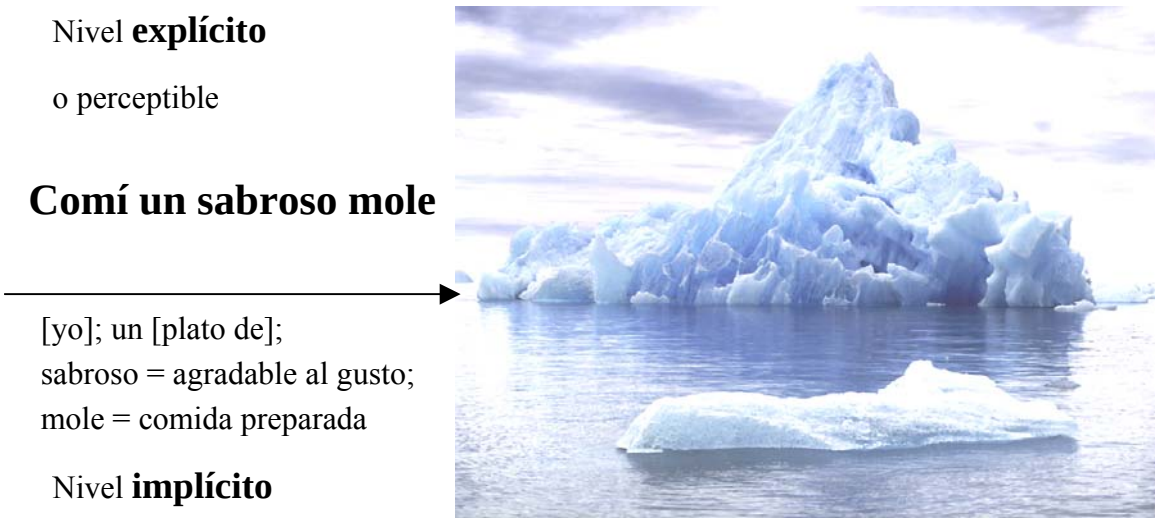

Figura 1 El contexto

Pero, ¿Qué hay "debajo del agua"? Debajo del nivel del agua (indicado con la flecha) existe la información **implícita** requerida para la interpretación completa, que depende de: el conjunto de reglas gramaticales; los vocabularios del emisor (hablante o escritor) y del receptor (oyente o lector) idealmente iguales; la habilidad del emisor para expresar sus ideas; la habilidad del receptor para integrar la información del discurso; la información previa o el conocimiento del tema; el conocimiento enciclopédico o del sentido común; y el acto comunicativo o actividad lingüística. *Se percibe sólo el texto* y lo demás permanece oculto; ***el témpano completo***

constituye el contexto del discurso necesario para el proceso de generación o interpretación del lenguaje natural.

De acuerdo a lo anterior, el contexto del discurso *está en función de la situación determinada que da origen o produce una actividad lingüística*. Si esta situación es bien conocida, por el emisor y el receptor, se requiere el uso pocas palabras y de muchas si se ignora totalmente; conforme al principio de minimización "*Entre menos se dice, mayor significado tiene (cada vocablo)*" [Levinson, 1987:68]. Hay así una relación inversa entre la comunicación y la situación, y esta relación *tiende a una constante informativa* [Cerdá, 1975] como se ilustra en la figura 2.

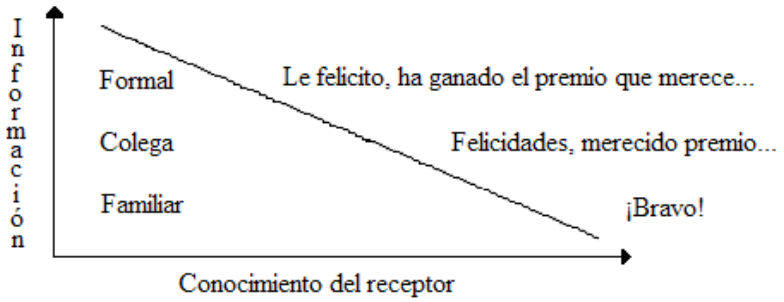

Figura 2 Principio de Minimización

Se puede observar que la situación, de felicitación por la entrega de un premio, depende de cuanta información haya sido intercambiada previamente por el emisor y el receptor (en conversaciones anteriores) y se supone que será mayor si existe una relación de confianza sobre las expectativas, esfuerzos por conseguir el premio y que regresa después de obtenerlo; así la relación empática y más completa con un familiar sólo requiere de un ¡bravo! ya que se conocen previamente los hechos. En cambio si la situación se presenta con el ejecutivo que entrega el premio, es necesario mencionar con mayor amplitud las razones de la felicitación. En algunos manuales de redacción [Schmelkes, 2002] se recomienda tomar en cuenta esta consideración cuando se escribe un documento, ya que es necesario ubicar la audiencia (posibles lectores), para dar la explicación más amplia y a la vez concisa.

2.3 El texto y sus propiedades

Aunque hoy en día no hay un gran consenso para definir el término texto, se considera que *es cualquier conjunto estructurado de enunciados que se producen en un proceso de comunicación.* Los textos pueden ser *orales o escritos*; literarios o no; para leer o escuchar, o para decir o escribir; largos o cortos; etc. Por lo tanto, se consideran textos los escritos de literatura, los diálogos y las conversaciones, las noticias, las pancartas publicitarias, etc.

Los enunciados que forman un texto están en función de lo que se quiere expresar: un enunciado puede contener información que amplíe, explique, corrija o contraste lo dicho anteriormente. Para que una manifestación verbal pueda ser considerada como texto deberá cumplir al menos las propiedades textuales de: adecuación, cohesión y coherencia. La adecuación va íntimamente ligada al contexto sociocultural del discurso; la cohesión a la dependencia gramatical entre las diferentes unidades que lo componen; y la coherencia al tipo de conocimiento que se requiere para interpretar el texto.

2.3.1 La adecuación

La adecuación es la propiedad que *considera el conocimiento y el dominio de la diversidad lingüística*. La lengua no es uniforme ni homogénea, sino que presenta variaciones según diversos factores: la geografía, la historia, el grupo social, la situación de comunicación, la interrelación entre los hablantes, el canal de comunicación, etc.

Dentro de un mismo dialecto, la lengua también ofrece registros muy diferentes: especializados, formales, coloquiales, etc. Por ejemplo *gastralgia, dolor de estómago*, *dolor de panza* o *dolor de barriga* pueden ser sinónimos en algunos contextos socioculturales, pero tienen valores sociolingüísticos diferentes: *gastralgia* está marcada formalmente y pertenece a un registro más culto y especializado como es la medicina (*gastro* = estómago y *algia* = dolor) que se usa en reportes de hospitales; *dolor de estómago* pertenece a un nivel de formalidad familiar o neutra, se utiliza en general y hasta en la publicidad comercial de radio y televisión; *dolor de panza* o *dolor de barriga* se considera coloquial si lo dice un niño en el ambiente familiar, o vulgar si lo expresa un adulto en público.

Un texto adecuado permite suponer que el autor supo escoger de entre todas las soluciones lingüísticas que da la lengua, la más apropiada para cada situación de comunicación. Para ello, es necesario utilizar el dialecto local o el estándar más general según los casos; y también es necesario dominar cada uno de los registros más habituales de la lengua: los medianamente formales, los coloquiales, los especializados más utilizados por el hablante, etc. Esto implica tener bastante conocimiento, aunque sea subconsciente, sobre la diversidad lingüística de la lengua: saber qué palabras son dialectismos locales, y que por lo tanto no serían entendidas fuera de su ámbito, y cuáles son generales; conocer la terminología específica de cada campo. En resumen, la adecuación *exige al usuario* de la lengua *sensibilidad sociolingüística para seleccionar el lenguaje apropiado en cada comunicación*.

2.3.2 La cohesión

La cohesión *es la propiedad que une el texto tomando en cuenta las articulaciones gramaticales*. Las oraciones que conforman un texto no son unidades aisladas e inconexas, puestas una al lado de otra, sino que están vinculadas o relacionadas con medios gramaticales diversos (puntuación, conjunciones, artículos, pronombres, sinónimos, entonación, etc.), de manera que conforman entre sí una imbricada red de conexiones lingüísticas, la cual hace posible su codificación y decodificación. En pocas palabras, la propiedad de la cohesión engloba cualquier mecanismo de carácter lingüístico o paralingüístico que sirva para relacionar las frases de un texto entre sí; es básicamente gramatical y afecta a la formulación superficial del mismo.

Para dotar de cohesión al texto se conocen diferentes mecanismos (o sistemas de conexión de oraciones) y figuras de construcción que consisten en reglas de concordancia de caso, género, número y persona o en alteraciones del orden "*normal*" de la frase entre las que se incluyen: la referencia, la deixis, la anáfora, las relaciones temporales (tiempos verbales), las relaciones semánticas entre palabras, los mecanismos paralingüísticos, la entonación, la puntuación, el hipérbaton, el pleonasmo, la elipsis y la silepsis.

Cabe mencionar que la entonación y la puntuación se consideran dos sistemas de cohesión paralelos con características y funciones particulares. *La entonación* es uno de los más importantes y expresivos mecanismos de cohesión que *sólo se da en la lengua oral*; en contraste,

la puntuación es propia de la escritura con las posibilidades de expresión limitadas a los signos gráficos.

La entonación indica si una oración termina o no, si se ha acabado de hablar, o si se trata de una interrogación, una admiración o una afirmación, etc.; tiene también otras funciones y capacidades expresivas que van mucho más allá de la cohesión: indica la actitud del hablante (seria, irónica, dubitativa, reflexiva, etc.), el énfasis que se pone en determinados elementos del texto: una palabra, una frase, etc. Si bien es cierto que determinadas formas de entonación se marcan en el escrito con signos gráficos apropiados (**? ,! , -**), otros muchos usos de la puntuación (oposiciones, enumeraciones, cambios de orden, etc.) tienen una explicación únicamente sintáctica, sin correlación tonal.

2.3.3 La coherencia

La coherencia es la propiedad que *indica cuál es la información pertinente que se ha de comunicar y cómo se ha de hacer* (en qué orden, con qué grado de precisión o detalle, con qué estructura, etc.). En pocas palabras, la coherencia es la propiedad que *se encarga de la cantidad, la calidad y la estructuración de la información*; es básicamente semántica y afecta a la organización profunda del significado del texto.

Entre el conjunto de medios que existen para conseguir la coherencia textual se tienen:

1. **Las presuposiciones.**- Se trata de la información que el emisor del texto supone que conoce el receptor. Es esencial para que un texto sea coherente, para el receptor, que el emisor haya "acertado" en sus presuposiciones.
2. **Las implicaciones.-** Se trata de las informaciones adicionales contenidas en un enunciado. Un enunciado del tipo "*cierra la puerta*" contiene, al menos, tres implicaciones: hay una puerta, la puerta está abierta y el receptor está en condiciones de cerrarla.
3. **El conocimiento del mundo.**- La coherencia de un texto depende también del conocimiento general que se tenga del mundo. Por ejemplo, un enunciado del tipo "*Los cuervos están de luto*" contradice el conocimiento general "normal" de la realidad porque los cuervos son considerados aves de carroña (comen cadáveres) y no lamentan la muerte de un ser viviente.

4. **El marco referencial** - Se trata del tipo de texto, su finalidad y la situación comunicativa en la que se produce. Dependiendo del marco, un determinado enunciado puede ser coherente, aunque choque con el conocimiento general "normal" del mundo. Por ejemplo, el enunciado considerado anteriormente, "*Los cuervos están de luto*", sería coherente en un texto que trate de películas mexicanas ya que es el título de una de ellas.
5. **Tema y Rema** - El **tema** es el asunto o materia del discurso; es aquello de lo que se habla o escribe y a lo que se deben subordinar todos y cada uno de los enunciados del texto; es lo que el emisor supone "*conocido*" por el receptor y sirve de base para recibir lo "*desconocido*" o nueva información que se denomina **rema** o comentario. El equilibrio entre lo que ya se sabe y lo desconocido asegura la comprensión y el interés de la comunicación y sólo cuando esta correlación tema-rema se ajusta adecuadamente la comunicación tiene éxito. Además, el tema y el rema van cambiando a medida que el receptor decodifica el texto, porque lo que es desconocido ($rema_1$) pasa a ser conocido (o parte del tema) y sirve de puente para recibir los nuevos datos ($rema_2$, $rema_3$, ... $rema_n$). Este proceso se conoce como tematización y es la base de la generación o interpretación progresiva de la información en el texto.

Pero ¿Qué es la coherencia? En los diccionarios se le define como "*la conexión o unión de una cosa con otra*" [DRAE, 1995] y como "*conexión o enlace lógico de una cosa con otra*" [ESPASA, 2001]. Aplicada al discurso se puede definir como "*la conexión, continuidad o coordinación, que se observa entre los componentes del discurso*". Se han identificado cinco tipos de coherencia, que no son independientes entre sí y se encuentran íntimamente interrelacionadas: **temporal**, la que identifica la continuidad de ***cuando*** los hechos están ocurriendo; **localidad**, que identifica el ***luga***r donde ocurren los eventos; **causal**, que consiste en ***el porqué*** (las razones) los hechos ocurren; **estructural**, que tiene que ver con ***la forma*** en que se describen los hechos en el discurso [Gernsbacher, 1997]; y **referencial**, la que identifica a ***quién o qué*** se está discutiendo.

La coherencia referencial, íntimamente ligada a la anáfora indirecta, se observa como un proceso incremental de procesamiento de información transmitida del emisor (escritor o hablante) al receptor (lector u oyente) tanto en lenguaje escrito como en el oral [Gernsbacher, 1997]. Este

proceso tiene que ver con señales que el emisor coloca explícitamente en el texto o discurso; se requiere pues, el reconocimiento de estas señales por el receptor para interpretar la coherencia de la nueva información con la previamente recibida. En el texto se pueden identificar señales que en forma explícita hacen referencia a entidades mencionadas previamente en el discurso; por ejemplo en la anáfora los pronombres él, ella, etc.

(5) **María** terminó su noviazgo con **Juan**. *Él* se molestó mucho con *ella*.

Otras señales no están explícitamente en el texto y llegar a identificar las entidades a que hacen referencia requiere un conocimiento del proceso; por ejemplo el fenómeno de elipsis (u omisión) del pronombre, "implícito" en la conjugación del verbo del siguiente ejemplo.

(6) Ayer [*tu*] jugaste un buen partido de fútbol.

Algunas señales de coherencia están aún más ocultas e identificar las entidades a que hacen referencia requiere un proceso de inferencia. Para interpretar estas señales el receptor debe apoyarse en su conocimiento previo adquirido del mundo real (de los eventos, hechos y relaciones). Ejemplo:

(7) Juan N. fue **asaltado** ayer. *El ladrón* continúa prófugo.

En la nota periodística anterior se observa que la coherencia requiere conocimiento previo de que en un asalto (evento) participan la víctima del asalto (Juan) y el ladrón o asaltante.

La estrategia de solución intentada para resolver la coherencia textual, es resolver cada tipo de coherencia uno por uno, por lo que este trabajo se enfoca sólo a la ***coherencia referencial*** considerando que la anáfora indirecta es un tipo de referencia que se apoya en las relaciones entre entidades del discurso para interpretar el texto.

2.4 Trabajo relacionado

Se encontraron pocos trabajos realizados, desde el punto de vista de la lingüística computacional, sobre la anáfora indirecta; de los encontrados, se consideran cuatro representativos: dos dedicados al Japonés [Murata, 1996, 2000] y dos al Inglés [Gelbukh y Sidorov, 1999; Muñoz et al, 2000], ninguno al Español. Una es la tesis "Resolución de la anáfora en oraciones del japonés usando expresiones superficiales y ejemplos" [Murata, 1996] sobre la resolución de la anáfora en general con el capítulo 4 dedicado a la anáfora indirecta en particular y un artículo [Gelbukh y Sidorov, 1999] donde se propone un método de resolución de la anáfora indirecta.

En la tesis de Murata, se propone un método, basado en el modelo del tópico o focal, para resolver la anáfora indirecta en el Japonés utilizando las relaciones existentes entre dos verbos, almacenadas en un diccionario de marcos basado en casos típicos. Primero toma todos los posibles antecedentes del tópico o foco de las oraciones precedentes; en segundo lugar, pondera dichos antecedentes de acuerdo a su plausibilidad; y por ultimo, determina el antecedente requerido combinando la ponderación de los antecedentes, el peso de la similaridad semántica de cada relación almacenada en el diccionario y el peso relativo de la distancia entre la anáfora y su posible antecedente. Obtuvo una precisión de 68% y una especificidad (recall) de 63% en las oraciones de prueba comprobando que el uso de las relaciones es útil.

En el artículo, de Gelbukh y Sidorov [1999], el método detecta la anáfora indirecta expresada con los marcadores más frecuentes, en el Inglés e identificados por ellos, un artículo definido o un demostrativo y aplican el modelo de escenario basado en diccionario.

Se utiliza para descubrir relaciones anafóricas entre palabras en diferentes oraciones "*entre una palabra y una entidad implícitamente introducida en el texto previo*"; dicha entidad no tiene una representación superficial en el texto sino en el escenario prototípico de la palabra antecedente. Utilizan un diccionario donde cada "entrada" de palabra está relacionada con las palabras que pueden participar potencialmente con la situación expresada por la "entrada". Establecen, en el ámbito sintáctico, dos condiciones que hacen posible la presencia de la anáfora indirecta como condiciones necesarias (pero no suficientes); una vez detectada la anáfora potencial se buscan los posibles candidatos para antecedentes con base en la distancia lineal y

estructural; se determina el grado de satisfacción por conteo hasta lograr un nivel de satisfacción preestablecido. Si se logra, significa que existe la relación anafórica indirecta de otra forma se supone inexistente.

El algoritmo de Gelbukh y Sidorov [1999a y 1999b] no prueba las palabras dentro de la misma frase simple y para la resolución de la anáfora indirecta toma en cuenta los dos problemas fundamentales: descubrir la presencia de la anáfora indirecta y resolver la ambigüedad de la relación anafórica.

El acercamiento al problema se hace en el orden opuesto; se intenta resolver la relación anafórica, y si se tiene éxito, se considera que el elemento de esta relación se encuentra en el discurso. El algoritmo para detectar la anáfora trabaja como sigue:

- Se considera cada palabra del texto.
- Si la palabra está precedida por un artículo determinado o un pronombre demostrativo es una anáfora potencial y el algoritmo intenta encontrar un antecedente plausible para él, buscando los posibles antecedentes candidatos con base en la distancia lineal y estructural de la anáfora potencial.
- Para cada antecedente potencial, se prueban las condiciones de referencia implícita y grado de compatibilidad. El grado de satisfacción, en lugar de una respuesta binaria de sí o no, se determina como una probabilidad; así, se combinan (multiplican) las probabilidades para las condiciones y la distancia, y se utiliza un valor límite para decidir cual pareja de palabras pasa la prueba, lo que significa que se encontró una relación anafórica o no dependiendo del resultado.
- El algoritmo se detiene cuando encuentra el fin de archivo (no hay más palabras por revisar).