TEORÍAS/MODELOS LINGÜÍSTICOS QUE SE PUEDEN APLICAR A LA TAREA DE GENERACIÓN AUTOMÁTICA DE RESÚMENES*

Elena Lloret

31 de octubre de 2008

1. INTRODUCCIÓN A LA LINGÜÍSTICA DEL TEXTO

Según [1], cualquier **texto** ha de cumplir siete criterios de textualidad y tres principios básicos que regulan la comunicación textual. Los criterios de textualidad pueden ser de tipo lingüístico, psicolingüístico, sociolingüísitco o computacional, y son los siguientes:

- lingüísticos:
  - cohesión
  - coherencia
- psicolingüísticos:
  - intencionalidad
  - aceptabilidad
- sociolingüísticos:
  - situacionalidad
  - intertextualidad
- computacional:
  - informatividad

Los tres principios comunicativos son *eficacia* (obtener mejores resultados comunicativos invirtiendo un esfuerzo mínimo), *efectividad* (está relacionada con el impacto comunicativo que el texto provoca en el receptor) y *adecuación* (es el equilibrio óptimo que se consigue en un texto entre el grado de actualización de los criterios de textualidad y de las demandas comunicativas).

En cuanto a los criterios de textualidad, comentados de uno en uno, tenemos que la *cohesión* consiste en que las secuencias de las oraciones que componen la superficie textual están interconectadas a través de relaciones gramaticales, como la *repetición*, las *formas pronominales*, la *correferencia*, la *elisión* o la *conexión*. Por otra parte, respecto a la *coherencia* diremos que un texto posee coherencia cuando los conceptos que componen su universo del discurso están interconectados a través de relaciones de diversa naturaleza, como por ejemplo, la relación de causalidad). La *intencionalidad*, sin embargo, consiste en que la organización cohesiva y coherente del texto sigue un plan dirigido hacia el cumplimiento de una meta. La *aceptabilidad*

se manifiesta cuando un receptor reconoce que una secuencia de enunciados constituye un texto cohesionado, coherente e intencionado porque lo que se transmite en el texto es, según su opinión, relevante. La *situacionalidad* se refiere a los factores que hacen que un texto sea pertinente en un determinado contexto, mientras que la *intertextualidad* se refiere al hecho de que la interpretación de un texto depende del conocimiento que se tiene de textos anteriores. Finalmente, la *informatividad* es el factor de novedad que motiva el interés por la recepción de un texto.

Estas normas de textualidad tienen que ver con la manera en qué se conectan unos elementos con otros: mediante dependencias gramaticales en la superficie (cohesión); mediante dependencias conceptuales del mundo textual (coherencia); mediante las actitudes de los interlocutores hacia el texto (intencionalidad y aceptabilidad); mediante la incorporación de lo nuevo en lo ya conocido (informatividad); mediante la adecuación a la situación (situacionalidad) y mediante la interpretación de la dependencia entre textos distintos (intertextualidad).

Cuando un receptor recibe un texto, éste intenta emular el proceso de producción del mismo, intentando recuperar las ideas y los planes principales que subyacen en dicho texto. A continuación, se expondrán de forma más detallada los criterios de textualidad relativos a la cohesión, la coherencia, la informatividad y la intertextualidad.

1.1. Cohesión

La estabilidad de un texto se mantiene gracias a la continuidad de los elementos que lo integran. La noción de continuidad se basa en suponer que existe una relación entre los diferentes elementos lingüísticos que configuran el texto.

Las unidades sintácticas principales son el **sintagma** (compuesto po un núcleo modificado al menos por un elemento), la **cláusula** (unidad compuesta al menos por un sustantivo o un sintagma nominal, que concuerda con un verbo o sintagma verbal) y la **oración** (una unidad compuesta, al menos por una cláusula dependiente). Por otro lado, existen una serie de **mecanismos** que contribuyen a la **continuidad de un texto** y a economizar, a su vez, el esfuerzo de procesamiento. Estos mecanismos son los siguientes: *repetición* (reutilización directa de elementos o de patrones idénticos), *repetición parcial* (permite utilizar el mismo elemento utilizado con anterioridad pero con diferente categoría y no exactamente igual, por ejemplo: andar – los andares), *paralelismo* (repetición de una estructura enriquecida por la aportación de nuevos elementos), *paráfrasis* (repetición de un mismo contenido, pero transmitido mediante expresiones lingüísticas distintas), uso

de *proformas* (permite reemplazar elementos independientes portadores de contenido por formas dependientes más breves, por ejemplo "se"), *elisión* (repetición incompleta de una estructura en la que se ha omitido alguna expresión), *tiempo y aspecto verbales* (permite marcar relaciones existentes entre los elementos lingüísticos), *conexión* (uso de conectores para insertar señales superficiales) y *entonación* (en la comunicación oral, para marcas la importancia o novedad de un contenido).

Desde el punto de vista de la teoría lingüística de Van Dijk [7], el sintagma, la clásula o la oración serían *macroestados* gramaticales, mientras que los elementos que los componen serían *microestados* del sistema textual. En la sección 4 se introducirán los conceptos de macro y microestado.

1.2. Coherencia

El sentido de un texto hace referencia al conocimiento que se transmite de manera efectiva mediante las expresiones lingüísticas que aparecen en el texto. La continuidad en el sentido de un texto constituye la base de la coherencia. La coherencia también puede entenderse como el resultado de la combinación de los conceptos y de las relaciones en una red compuesta por espacios de conocimiento que giran alrededor de los temas principales. Los candidatos más adecuados para ejercer de centros de control se denominan **conceptos primarios** y están constituidos por:

- OBJETOS: entidades conceptuales con una identidad y una constitución estables.
- SITUACIONES: configuraciones de objetos presentes en sus estado habituales.
- ACONTECIMIENTOS: elementos que modifican una situación o un estado dentro de una situación.
- ACCIONES: acontecimientos realizados intencionalmente por un agente.

La hipótesis sintáctica de que la función de núcleo gramatical la desempeñan normalmente conceptos primarios se confirma en tantas ocasiones en los textos como para que se pueda generalizar como afirmación discursiva[1]. Por otra parte, los modificadores gramaticales funcionan como atributos, localizaciones, estados, etc., denominándose **conceptos secundarios**. En relación a éstos, encontramos:

- ESTADO: condición temporal de una entidad.

- AGENTE: fuerza que posee una entidad que realiza una acción, y que de esa manera, modifica una situación.
- ENTIDAD AFECTADA: la entidad cuya identificación se ve modificada por un acontecimiento o una acción en la que no figura ni como agente ni como instrumento.
- RELACIÓN: relaciones de tipo "padre-hijo", "jefe-empleado".
- ATRIBUTO: condición característica de una entidad.
- LOCALIZACIÓN: posición espacial de una entidad.
- TIEMPO: posición temporal de una situación, estado o acontecimiento.
- MOVIMIENTO: cambio de localización.
- INSTRUMENTO: objeto que proporciona los medios para que suceda un acontecimiento.
- FORMA: configuración, contorno, parecido.
- PARTE: un componente de una entidad.
- SUSTANCIA: materiales de los que se compone una entidad.
- CONTENCIÓN: localización de una entidad dentro de otra pero no como parte o sustancia.
- CAUSA: manera en que una situación o acontecimiento influye en las condiciones que han de darse para que ocurra otro acontecimiento.
- POSIBILIDAD: manera en que una situación posibilita (crea las condiciones suficientes, pero no necesarias) para que se lleve a cabo otra.
- RAZÓN: cuando una acción es el resultado esperable de un acontecimiento previo.
- PROPÓSITO: cuando se planea intencionadamente que suceda un acontecimiento a partir de otro.
- PERCEPCIÓN: operaciones de entidades creadas sensorialmente durante las que el conocimiento se integra mediante los órganos sensoriales.
- COGNICIÓN: almacenamiento, organización y utilización del conocimientos mediante entidades creadas sensorialmente.

- EMOCIÓN: estado experiencial y no neutral con respecto a una entidad creada sensorialmente.
- VOLICIÓN: actividad de desear mendiante una entidad creada sensorialmente.
- RECONOCIMIENTO: emparejamiento exitoso entre percepción y conocimiento previo.
- POSESIÓN: relación en la que una entidad creada sensorialmente se cree predominante y controla una entidad.
- EJEMPLO: miembro de una clase que hereda todos los rasgos no cancelados de la clase a la que pertenece.
- ESPECIFICACIÓN: relación entre una superclase y una subclase, con una afirmación de lso rasgos más restringidos de la última.
- CANTIDAD: concepto numérico, de alcance, escala o medida.
- MODALIDAD: concepto de necesidad, probabilidad, posibilidad, permisibilidad, obligación o sus opuestos.
- SIGNIFICANCIA: un significado simbólico asignado a una entidad.
- VALOR: asignación del equivalente de una entidad en términos de otras entidades.
- EQUIVALENCIA: igualdad, correspondencia y parecido.
- OPOSICIÓN: contrario a equivalencia.
- CORREFERENCIA: relación en la que expresiones diferentes activan la misma entidad del mundo textual.
- REPETICIÓN: relación en que la misma expresión reactiva un concepto, pero no necesariamente con la misma referencia a una entidad o con el mismo sentido.

Atendiendo a estos dos tipos de conceptos, podríamos construir un *espacio de conocimiento* coherente para el texto, es decir, un *macroestado* conceptual donde los conceptos son *microestados*. Pero para hacer coherente toda la información, el receptor debe realizar inferencias, esto es, suplementar los conceptos y las relaciones que se manifiestan en el texto con el fin de rellenar los huecos o discontinuidades que aparezcan en el mundo textual.

1.3. Informatividad

El concepto de informatividad se relaciona con el grado de novedad o de imprevisibilidad que tiene un texto para sus receptores. El nivel de informatividad de un texto se valora en función de su contenido. La importancia del contenido se explica por el papel dominante que juega la coherencia en la textualidad. Existen tres niveles de informatividad: nivel superior, nivel inferior y nivel extraordinario. Las palabras funcionales (artículos, preposiciones y conjunciones), es decir, todas aquellas palabras que marcan relaciones y que carecen de contenido, aparecen con tanta frecuencia en cualquier texto, que su presencia pasa normalmente inadvertida. Las palabras con contenido, sin embargo, son más informativas que las funcionales. Cuanto más previsible sean los elementos contenidos en el texto, menos informativo será pero más fácil será para el receptor procesarlo. Sin embargo, en un tercer nivel de informatividad formado por elementos infrecuentes, requerirá el empleo de abundantes recursos de procesamiento, aunque la información contenida en el texto, será más interesante tanto para el productor como para el receptor. Una de las marcas formales que hacen evidente la existencia de expectativas por parte de los interlocutores es el *grado de definición* de los artículos que aparecen en el texto (artículos definidos frente a indefinidos). Las entidades definidas en el mundo textual serán identificables, accesibles y recuperables sin un excesivo esfuerzo de procesamiento. Por el contrario, el artículo indefinido se reserva para entidades que aparecen por primera vez en el discurso. Esta utilización de la indefinición y definición, da lugar a lo que se denomina acceso procedimental. Si una entidad textual se encuentra activada en la memoria inmediata, entonces se propicia que el acceso al **conocimiento definido y prototípico** de la misma sea fácil y rápido. En contraste a esto, una entidad acompañada por un artículo indefinido consideraremos que es un **miembro accidental de una categoría o clase**. Aparte de la definición/indefinición, otra de las marcas formales es la ordenación secuencial (secuenciación) de las oraciones o cláusulas que aparecen en el texto, que se organizan mediante el uso de conjunciones.

La coherencia textual se basa en la explotación discursiva de las expectativas que genera la información que va apareciendo progresivamente en el texto.

1.4. Intertextualidad

El término intertextualidad se refiere a la relación de dependencia que se establece entre, por un lado, los procesos de producción y de recepción de un texto determinado, y por otro, el conocimiento que tengan los participantes

de otros textos anteriores relacionados con él.

Existen diferentes tipos de textos, dando lugar a una amplia tipología textual. Por ejemplo, en los **textos descriptivos**, los centros de control son las situaciones o los objetos, y se da una elevada frecuencia de aparición de relaciones conceptuales de atribución de características, de estados, de ejemplificación y de especificación. La superficie textual de estos textos reflejará una elevada densidad de modificadores y complementos. El patrón que se suele aplicar es el marco. Por el contrario, los **textos narrativos** se utilizan para organizar discursivamente las acciones y los acontecimientos en un orden secuencial determinado. En este tipo de texto abundarán relaciones conceptuales para marcar la causa, la razón, el propósito, la posibilidad, y la proximidad. La superficie textual reflejará una elevada densidad de estructuras subordinadas y el patrón habitualmente más aplicado para esta clase de textos es el esquema. Los **textos argumentativos**, por otra parte, se utilizan para persuadir al receptor textual de que determinadas creencias o ideas son verdaderas o falsas, favorables o desfavorables para sus intereses. En este tipo textual aparecen con mucha frecuencia relaciones conceptuales para expresar la razón, la significación, la volición, el valor y la oposición. La superficie textual argumentativa reflejará una elevada densidad de mecanismos cohesivos que expresan el énfasis y la insistencia, por ejemplo, la repetición, el paralelismo o la paráfrais. El patrón que se aplica para este tipo de textos es el plan.

Existen tres tipos de perspectivas teóricas distintas sobre la rememoración de materiales que guardan similitud con el contenido informativo de los textos: *abstracción de huellas*, *construcción* y la *reorganización* (reconstrucción).

2. Principio de la cantidad de la codificación

Para establecer las relaciones entre las partes de un texto, las lenguas ofrecen varios recursos, puesto que han sido moldeadas para satisfacer la necesidad humana de comunicación. Una noción importante al respecto es la de la "coherencia informativa" [4]. No sólo el discurso constituye un compromiso informativo. Cada cláusula dentro de un discurso coherente y cohesivo debe contener igualmente algo de información dada para conectarla con lo anteriormente dicho, y también algo de información nueva para no ser redundante. Según Givón [4], una cláusula contiene sólo un ítem de información nueva. Los otros elementos están establecidos y tienen la función de integrar la información nueva. Son indispensables para procesarla en un tiempo normal y hacer comprensible el discurso. Para presentar la información nueva

y dada en la cláusula y en el discurso, en [5] se establece el **Principio de la cantidad de la codificación** (the code quantity principle), que dice así: *"mientras menos predecible, o más importante, es la información, más prominente, más evidente y larga será el medio de codificación que la represente"*. Esto quiere decir que si la información es más relevante se codificará con mayor peso léxico. Para ello, la lengua provee los medios para que esto se cumpla. Así, existe una unidad que es capaz cargar más o menos información según la necesidad: el **sintagma nominal**. Éste acepta modificadores y extensiones que los permite aparecer con mayor peso. Otra forma de aumentar el peso del sintagma nominal es mediante una clásula relativa. Ni los verbos ni los adverbios tienen esta misma característica, ya que un grupo verbal contiene solamente un elemento lexical (el verbo), y los grupos adverbiales tienen un alcance muy limitado [2]. Existe, pues, una relación proporcional entre la relevancia de la información y la cantidad de su codificación. Así, el sintagma nominal puede realizarse como un sintagma nominal léxico pesado o liviano, un pronombre o un elemento cero, de acuerdo al tipo de información que presenta. El principio de la cantidad de código tiene su fundamento cognoscitivo en el principio de la cantidad, la atención y la memoria (code-quantity, attention and memory), en el sentido de que la codificación más prominente y distinta atraerá más la atención del receptor y como consecuencia, la información que atrae más atención es memorizada, almacenada y recuperada más eficientemente. La información nueva lleva, por lo general, el acento principal de la cláusula y se presenta con mayor peso léxico para que el oyente encuentre el referente. En [2] se comenta que existe una interacción entre información dada e información nueva, puesto que algo que no es identificable para el receptor difícilmente puede estar ya en su consciencia. Lo que es identificable, lo es porque fue introducido anteriormente en el contexto lingüístico. Esta relación entre información nueva frente a información dada introduce los conceptos *tema* (información dada) y *rema* (información nueva) que se verán en la siguiente sección (sección 3).

3. Tema y Rema

El **tema** es el núcleo del texto, aquello de lo que se habla. Constituye el elemento con el cual el resto de la cláusula está conectado y que establece la relevancia de esta en el contexto, es decir, la manera como se conecta con las demás y sigue o cambia la línea textual. La continuidad en el tema contribuye a la coherencia del texto. Puede que se trate del mismo referente, el mismo lugar, misma acción, mismo objetivo, etc.

El **rema**, por el contrario, expresa lo nuevo, lo que se comunica acerca

del tema. Resulta más rico en información respecto al tema. La información del texto progresa a medida que se avanza en el rema. Cuando la información no tiene conexión con el tema o rema anterior, decimos que es incoherente, mientras que si repite la misma idea expresada ya en el tema estamos ante el caso de información redundante.

Finalmente, la idea aparecerá como resultado de la relación entre tema y rema.

3.1. Progresión temática

La estructura temática opera dentro de los límites de la oración (cada oración tiene un tema y rema). Estos permiten que el texto progrese y al mismo tiempo, conectan las oraciones entre sí y con el texto completo. En el momento de establecer el elemento que constituye el tema y el/los elementos que constituyen el rema en una determinada oración, hay que mirar más allá de ésta, y hay que considerar el texto en el que está insertada. Pero no hace falta tomar todo el texto, sino que con tomar una parte del mismo es suficiente. Se trataría de tomar como unidad más amplia el párrafo dónde se ubica la oración que estamos tratando, puesto que generalmente un párrafo marca una ruptura temática respecto al anterior.

En [3] se sugieren cuatro tipos de progresión textual que, a su vez, se relacionan con ciertos tipos de textos. Estos tipos son los siguientes:

- **Progresión lineal**: cuando el rema de una oración se convierte en el tema de la siguiente. Es caracterítica en los textos narrativos y descriptivos
- **Progresión constante**: el tema de la oración es constante y se repite como tema de la oración/es siguiente/es. Aparece en los textos descriptivos y expositivos.
- **Progresión derivada**: se da un hipertema que se divide en varios subtemas. Se da frecuentemente en textos expositivos.
- **Progresión convergente**: el tema resulta de la suma de dos o más ideas expresadas anteriomente. Aparece en todos los tipos de textos, pero sobre todo, en los conclusivos.

En las figuras 1 y 2 podemos ver los esquemas correspondientes al tipo de progresión temática constante y lineal, respectivamente, que son los dos tipos de progresiones temáticas básicos [6].

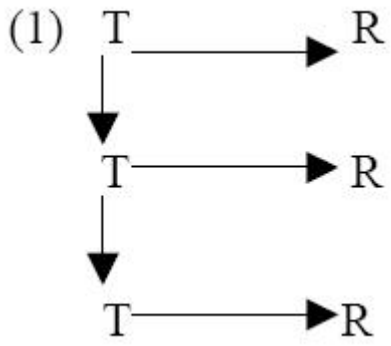

Figura 1: Progresión constante.

Cada unidad de información contiene como mínimo un foco (palabra o palabras destacadas). En inglés, la estructura temática se manifiesta mediante la posición de los constituyentes en la cláusula. El tema siempre aparece en primer lugar [2]. Ahora, definiremos un párrafo como un segmento de texto coherente centrado en un solo tema que comprende una única progresión temática, en la que el tema de la primera cláusula representa el tema del párrafo.

4. Estructura del texto

En secciones anteriores se han mencionado los conceptos de macroestructura y microestructura. Se puede hacer una distinción entre la microestructura y la macroestructura de un texto. La **microestructura** representa las relaciones entre las oraciones, mientras que la **macroestructura** representa las relaciones entre bloques de oraciones y la organización global de un texto. Considerando ambos niveles, estas *relaciones pueden ser temáticas o semánticas*. Las relaciones temáticas incluyen la anáfora, la correferencia y la "tematización". Entre las relaciones semánticas se incluyen las de causa-efecto, relaciones temporales y las relaciones lógicas [8].

4.1. La microestructura

Las relaciones que podemos encontrar en la microestructura de un texto incluyen la *anáfora*, los *enlaces semánticos*, la *cohesión semántica* y la *progresión temática*.

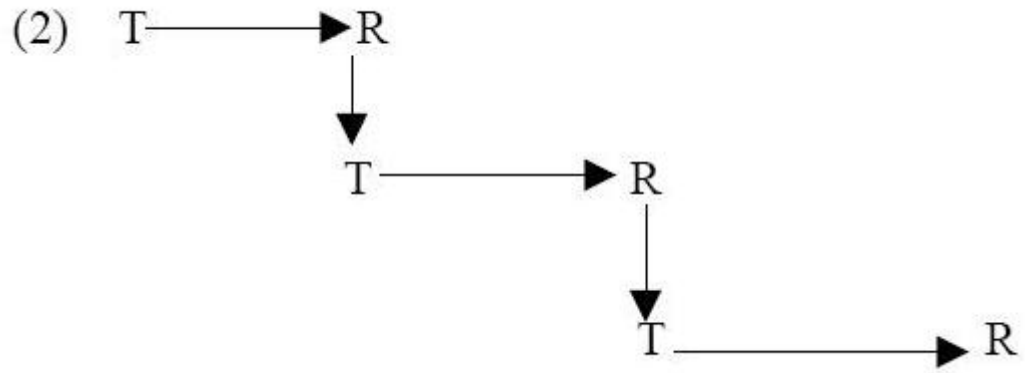

Figura 2: Progresión lineal.

El fenómeno de la **anáfora** es un fenómeno muy frecuente en la utilización de una lengua. Consiste en hacer referencia a algo que previamente ha aparecido en un texto. La primera vez que se nombra algo en un texto se suele hacer mediante expresiones indefinidas, mientras que en las posteriores referencias a los mismos se emplean expresiones definidas (pronombres, nombres genéricos, etc.). Estas relaciones pueden extenderse no sólo entre frases adyacentes, sino también a lo largo de todo el texto. Por otra parte, los **enlaces semánticos** se representan por la recurrencia de términos que pertencen al mismo campo semántico (por ejemplo, policía y ley). Las relaciones semánticas entre frases o cláusulas que aparecen seguidas en el texto, dotan a éste de cohesión local de tipo semántico. La progresión temática (véase sección 3.1) recoge la forma en la que la información nueva se expresa en el texto respecto a la información ya conocida.

4.2. La macroestructura

La macroestructura de un texto se deriva de las generalizaciones de las proposiciones que forman la microestructura y la eliminación de las cláusulas con información redundante o irrelevante. Dicha macroestructura se organiza atendiendo a *diferentes esquemas* que son inferidos por el receptor durante la comprensión del texto. El reconocimiento de dichas estructuras o esquemas por parte del receptor depende del conocimiento previo que éste tenga de las mismas. Ejemplos de macroestructuras son la macroestructura narrativa, o el esquema de los textos expositivos.

5. Producción de resúmenes desde el punto de vista lingüístico

Un resumen se puede considerar como una representación de la macroestructura de un texto hecha por un individuo teniendo en cuenta el conocimiento del mundo que posee. Es por esto, que el proceso de producción de resúmenes no es objetivo, sino subjetivo y es por esto por lo que no existe un resumen único para un texto. Teóricamente, en el **proceso de producción de un resumen** están envueltos **cuatro componentes**:

1. Comprensión de la microestructura.
2. Identificación del esquema (estructura) global del texto.
3. Aplicación de macro-reglas para generalizar y condensar la representación de la macroestructura.
4. Expresión de la macroestructura como un texto coherente.

Las macro-reglas que se proponen en [8] son la *"supresión"* (eliminación directa de contenido lingüístico), la *"generalización"* (reconsideración del material lingüísitico en un contexto más amplio) y la *"construcción"* (creación de material lingüístico nuevo). Estas macro-reglas se definen a nivel semántico, pero en el caso de la operación de supresión, ésta puede ser aplicada en la progresión temática para omitir o descartar las proposiciones que constituyen "remas" de las proposiciones "tema" .

Referencias

[1] Robert-Alain Beaugrande de and Wolfgang Ulrich dressler. *Introducción a la lingüística del texto.* Ariel, 1997.

[2] Annette Becker. Análisis de la estructura pragmática de la cláusula en el español de Mérida (Venezuela). *Revista de Estudios de Lingüística Española (ELiEs)*, 17, 2002.

[3] José M. Bustos Gisbert. *La construcción de textos en español.* Ediciones Universidad de Salamanca, 1996.

[4] Talmy Givón. *Syntax: A functional-typological introduction, I.* John Benjamins, 1984.

[5] Talmy Givón. *Syntax: A functional-typological introduction, II.* John Benjamins, 1990.

[6] John Hutchins. Summarization: Some problems and methods. In *Meaning: The Frontier of Informatics*, pages 151–173. Aslib, 1987.

[7] T.A Van Dijk. Complex semantic information processing. In *Walker, D.E. et al. (eds.) Natural language in information science*, pages 127–163, 1977.

[8] T.A Van Dijk. Recalling and summarizing complex discourse. In *Burghardt, W. and Holker, K. (eds.) Text processing*, pages 49–118, 1979.