CORRECTOR: UN SISTEMA DE VERIFICACION SINTACTICA Y ESTILISTICA DE TEXTOS

Consuelo Rodríguez Magro

Centro de Investigación UAM-IBM
Madrid
Centro de Tecnología de la Lengua
Sevilla

ducción

icio de este trabajo constituye un paso más en el desarrollo de herramientas utilizables rea de Procesamiento del Lenguaje Natural que, desde 19851 llevamos a cabo en el Investigación UAM-IBM [1]. [2]. [3]. [4], [5]. [6] y [7] Además, constituye en esta área: no existe ningún sistema de ayuda que detecte errores gramaticales y ilísticos en nuestra lengua. Hasta ahora, los editores de textos proporcionan una vel mucho más simple, como es la verificación y corrección ortográficas (p.e. **Word Perfect,** etc.). De todas formas, ha quedado patente que el uso de da a la escritura y elaboración de textos, en general, ha sido de gran utilidad. los usan, cada vez más, porque necesitan escribir muchos documentos. Como res que no se pueden detectar simplemente con un verificador ortográfico, iere mas información para hacerlo. Por eso, si, ademas de las ayudas que los nan hasta ahora, podemos contar con la posibilidad de una ayuda en la gramaticales y problemas estilísticos, las ventajas serán evidentes. Es lo el trabajo que presentamos.

estamos trabajando en el desarrollo de una gramática para el castellano, es CORRECTOR, un sistema de crítica textual en ordenador.

sistema de ayuda en la preparación de textos escritos. Informa sobre s y frases, diagnostica errores gramaticales, identifica posibles sugerencias y da explicaciones sobre el tipo de error en cuestión. gramática de amplia cobertura, SPG (Spanish PLNLP Grammar) extensa gramática que proporciona un análisis sintáctico para las tituye el rasgo principal que caracteriza a nuestro sistema.

y problemas de estilo.

distinción entre lo que entendemos por error gramatical y Podemos decir que existen diferencias de varios tipos:

1. **diferencia conceptual**

Los términos utilizados, error y problema, sugieren de qué tipo de diferencias se trata. Podemos decir que cuando se comete un error gramatical se están violando, claramente, las reglas gramaticales establecidas. Se trata de errores de **falta de concordancia, errores pronominales....**

Sin embargo, los problemas estilísticos pueden depender del tipo de texto y, en algunos casos, son más subjetivos. En general, no es que se viole ninguna regla gramatical, pero existe algún tipo de acuerdo en que algo ~suena mal~ o ~es preferible utilizar otro término~. El uso (abuso) de siglas o acrónimos, palabras extranjeras, etc. se encuadran en este grupo.

2. **diferencia de implementación**

Los errores gramaticales se identifican como una parte del análisis sintáctico, es decir, se especifican directamente en las reglas de la gramática, mientras que los problemas de estilo se identifican después del análisis sintáctico.

3. **diferencia de presentación al usuario**

Asimismo, la forma en la que el usuario ve el error cometido en nuestro sistema es diferente en un caso y en otro. Cuando se llama a CORRECTOR, tanto los errores gramaticales como los problemas estilísticos aparecen marcados, pero con un comentario diferente: los errores gramaticales con un asterisco y los problemas estilísticos con una barra vertical, ambos debajo del error. Los mensajes que aparecen también son distintos: para los errores gramaticales aparece la forma correcta por la que debe cambiarse el error, para los problemas de estilo sólo existe una descripción del problema, pero no aparece el sustituto.

No existen muchos libros que traten este tipo de cuestiones en nuestra lengua y que hagan un inventario claro de los errores que se cometen más frecuentemente es castellano. Las referencias más importantes con las que contamos son el **Libro de estilo** de EL PAIS [10], el **Diccionario de dudas** de M. Seco [11], el **Diccionario de uso** de M. Moliner [12] y algunas gramáticas españolas [13], [14] y [15]. No se trata, en ningún caso, de adoptar un tono dogmático, sino de comentar los errores que se cometen con más frecuencia y hacer sugerencias, como en el caso de las anomalías de estilo. Con los errores gramaticales, obviamente, se cometen otro tipo de transgresiones que deben evitarse siempre.

Errores y problemas que CORRECTOR trata

Entre los **errores gramaticales** considerados se encuentran:

- la falta de concordancia — concordancia nominal — concordancia verbal
- el fenómeno llamado "dequeísmo" *(Te prohibo de que vengas)*

• falta de preposición **de** cuando es necesaria, es el fenómeno opuesto al "dequeísmo" *(Estoy seguro que vendrá)*
• mala utilización del gerundio *(He recibido tu nota preguntando por los programas)*
• mala utilización de los pronombres, en posición enclítica o proclítica. Hablamos de casos de laísmo, leísmo o loísmo (La *he regalado un sombrero)*

Y entre los **problemas estilísticos:**

• signos de puntuación omitidos *(paréntesis, comas...)*
• uso erróneo de la puntuación, como la aparición de coma entre el sujeto y el verbo *(Las notas del tercer trimestre, se harán públicas mañana)*
• palabras repetidas
• acumulación de acrónimos
• utilización incorrecta de expresiones latinas (a *grosso modo)*
• utilización de palabras extranjeras *(meeting)* o traducción directa de las mismas *(chequear)*

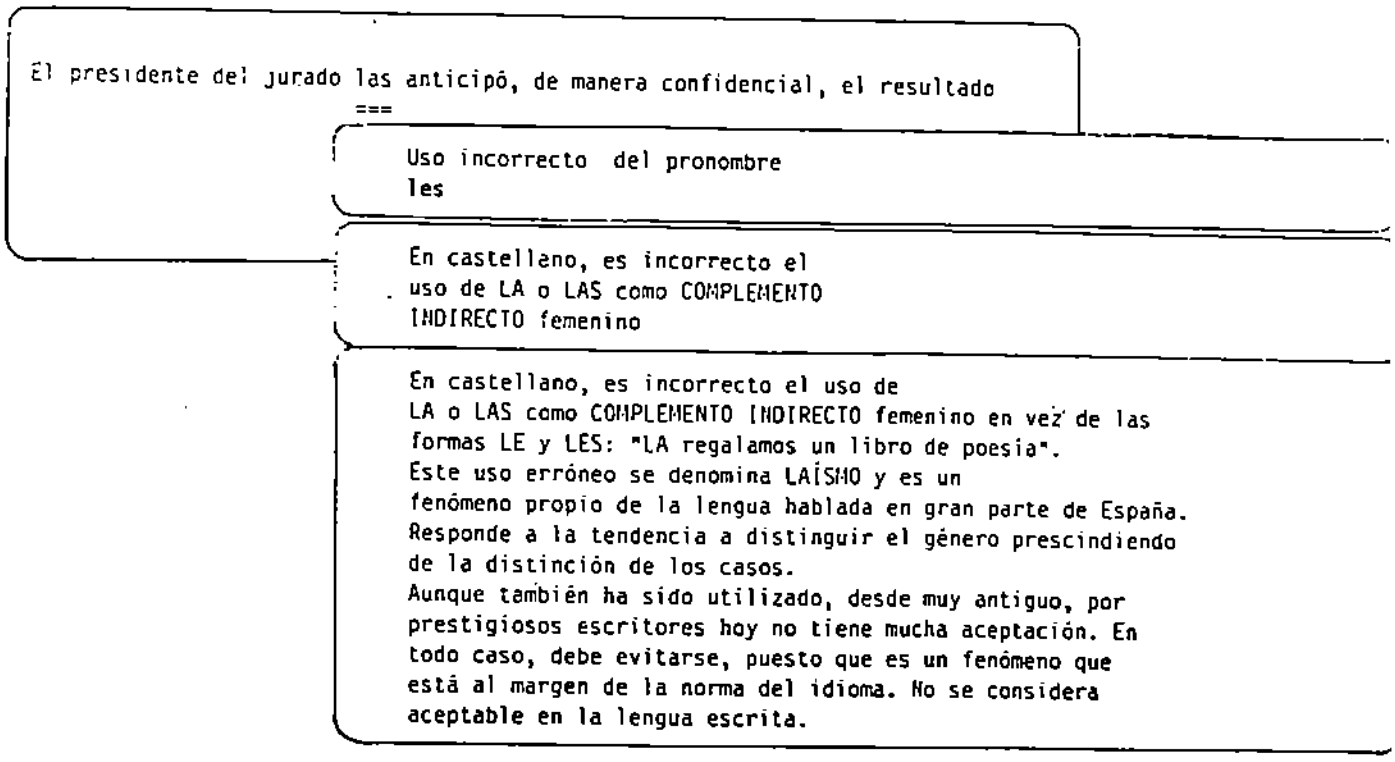
Ejemplificación de errores

Errores gramaticales

Aparecen marcados los errores en un texto y, debajo, las palabras por las que deben sustituirse:

```
El presidente del jurado las anticipó, de manera confidencial, el resultado
                            *les
del concurso.
La secretaria supuso de que la reunión era a las tres.
                         *que
```

CORRECTOR proporciona, a petición del usuario, tres niveles de información:



Asimismo, el usuario puede ver el árbol de análisis, si así lo desea:

```
¿Has chequeado el documento?
|ANGLICISMO
Las acciones que debemos tomar, se harán públicas mañana
                              |COMA INCORRECTA
```

Igualmente, se puede ver el árbol:

```
 Escriba una frase en castellano o un comando :
las acciones que debemos tomar, se harán públicas mañana.
-----------------------------------------------------------
DECL1  NP1      AJP1    ADJ1*   "las"
                NOUN1*  "acciones"
                VP1     NP2     PRON1*  "que"
                        VP2     VERB2*  "debemos"
                        VERB3*  "tomar"
       PUNC1    ","
       VP3*     NP2     PRON2*  "se"
                VERB4*  "harán"
                AJP1    ADJ1*   "públicas"
                AVP1    ADV1*   "mañana"
       PUNC2    "."
-----------------------------------------------------------

PROBLEMA DE ESTILO EN FRASE 12.
  SIGNO DE PUNTUACIÓN ERRÓNEO.
    las acciones que debemos tomar , se harán públicas mañana .
  CONSIDERE LA ELIMINACIÓN DE LA COMA
    las acciones que debemos tomar ? se harán públicas mañana  .
-----------------------------------------------------------

 Escriba una frase en castellano o un comando :
¿has chequeado el documento?
-----------------------------------------------------------
QUEST1 PUNC1    "¿"
       VERB1    "has"
       VERB2*   "chequeado"
       NP1      AJP1    ADJ1*   "el"
                NOUN1*  "documento"
       PUNC2    "?"
-----------------------------------------------------------
PROBLEMA DE ESTILO EN FRASE 17.
  USO INCORRECTO DEL VERBO (ANGLICISMO)
    ¿has chequeado el documento?
  INTENTE BUSCAR UN EQUIVALENTE EN CASTELLANO
-----------------------------------------------------------
```

Etapas en la detección de errores

Cuando se llama a **CORRECTOR** se siguen diferentes pasos en la detección de errores:

• primero se procesan las oraciones mediante un analizador léxico que está conectado a un diccionario con información morfológica y sintáctica

```
 Escriba una frase en castellano o un comando :
 el presidente del jurado las anticipó , de manera confidencial ,
el resultado del concurso .
-----------------------------------------------------------
DECL1  NP1*    AJP1    ADJ1*   "el"
                       NOUN1*  "presidente"
                ?      PP1     PREP1   "del"
                               NOUN2*  "jurado"
       NP2     PRON1*  "las"
       VERB1*  "anticipó"
       PUNC1   ","
       PP2     PREP2   "de"
               NOUN3*  "manera"
               AJP2    ADJ2*   "confidencial"
                       PUNC2   ","
       NP3     AJP3    ADJ3*   "el"
               NOUN4*  "resultado"
               PP3     PREP3   "del"
                       NOUN5*  "concurso"
       PUNC3   "."
ERROR GRAMATICAL EN FRASE 139.
  POSIBLE ERROR PRONOMINAL.
    el presidente del jurado las anticipó, de manera
   confidencial, el resultado del concurso .
  SUGERENCIA
    el presidente del jurado LES anticipó, de manera
   confidencial, el resultado del concurso .
-----------------------------------------------------------

 la secretaria supuso de que la reunión era a las tres .
-----------------------------------------------------------
DECL1  NP1     AJP1    ADJ1*   "la"
               NOUN1*  "secretaria"
       VERB1*  "supuso"
       SUBCL1  PREP1   "de"
               CONJ1   "que"
               NP2     AJP2    ADJ2*   "la"
                       NOUN2*  "reunión"
               VERB2*  "era"
        ?      PP1     PREP2   "a"
                       AJP3    ADJ3*   "las"
                       NOUN3*  "tres"
       PUNC1   "."
ERROR GRAMATICAL EN FRASE 145.
  POSIBLE ERROR PREPOSICIONAL
    la secretaria supuso de que la reunión era a las tres.
  SUGERENCIA
    la secretaria supuso QUE la reunión era a las tres.
-----------------------------------------------------------
```

Problemas de estilo

De la misma forma que en el caso de los errores gramaticales, el usuario queda informado del tipo de error.

• después la gramática analiza la oración

• si se viola una condición en una regla de la gramática y no se llega a producir un análisis, el sistema lo intenta de nuevo, relajando esa condición. Produce un análisis e informa del error, que será un **error gramatical**

• si se detecta un error después de completado el análisis, se trata de un **error estilístico**

El diccionario

Como acabamos de señalar, cada una de las palabras de una cadena se procesan con un analizador morfológico que está conectado a un diccionario con información morfológica y sintáctica. De esta forma, el diccionario devuelve para cada palabra toda la información. Veamos la información obtenida para alguna de las palabras que aparecen en el ejemplo anteriormente utilizado, "Las anticipó el resultado del concurso":

```
anticipó anticipar(VERB DITR (IND a) INDICAT) (OBJ s) PAST
                   PERS3 (RIND a) (RSPI a) SING (SUJ s) TR (PCODE SPV))

resultado resultar(VERB SCA CSADV (IND a) INTR (OBJ s)
                    PASTPART PINF QUASIAUX (SPS de) (SUJ (cq i s)) (PCODE SPV))
                    (NOUN MASC NPCOMP SING (PCODE SPMN))

concurso concursar(VERB INDICAT INTR (OBJ sp) PERS1 PRES
                    SING (SUJ s) TR (PCODE SPV))
                    (NOUN MASC SING (PCODE SPMN))
```

Una vez que el analizador ha procesado toda esta información, se manda al parser donde la gramática analiza su comportamiento sintáctico.

La gramática

La gramática está escrita en PLNLP (Programming Language for Natural Language Processing), lenguaje de programación especialmente diseñado para el procesamiento del lenguaje natural y está basado en las primeras investigaciones que George Heidorn realizó en 1972 **[13]**. PLNLP **[16]** se usa en el entorno LISP/VM en VM/SP.

PLNLP proporciona un entorno para escribir gramáticas que pueden generar (**encoding**) o analizar (**decoding**) oraciones. Nuestra gramática, SPG, de momento, sólo trabaja con la parte de análisis.

Las reglas de PLNLP se parecen a las reglas de estructura de frase, (APSG, Augmented Phrase Structure Grammar) [14]. Sin embargo, PLNLP no se basa en ninguna teoría gramatical determinada.

PLNLP manipula estructuras de datos conocidas como registros. y SPG crea una descripción sintáctica utilizando estos registros. Para que una regla determinada se aplique, deben encontrarse unas condiciones sobre unos determinados registros. Y cuando una regla se ejecuta, el resultado será la creación o la modificación de un registro.

Los registros son entidades que consisten en una colección de elementos llamados **atributos** y cada uno de estos atributos tiene un valor. A continuación podemos observar un ejemplo simplificado de registro:

```
STR       "el"
BASE      'EL'
POS       ' ADJ '
INDIC     SING MASC
```

PLNLP crea y modifica registros mediante reglas. Estas reglas tienen la siguiente forma:

```
ELEMENTO1(condiciones) ELEMENT02(condiciones)...
                          --> NUEVOELEMENTO(acciones)
```

Lo que aparece aquí como elementos se corresponde con los registros que PLNLP manipula (NP, VERB, SENT...). Las condiciones, los valores de ciertos atributos, se especifican en la parte izquierda de las reglas. Las acciones que deben tomarse se escriben en la parte derecha. Si los elementos de la izquierda coinciden con las categorías gramaticales de la frase que queremos analizar y si se cumplen todas las condiciones, se aplica la regla y se crea un nuevo elemento, el especificado en la parte derecha.

Veamos un ejemplo muy sencillo:

```
(3000)NP VP (NUMERO.AGREE.NUMERO(NP),PERSONA.AGREE.PERSONA(NP))
                                  --> VP(PRMODS=NP...PRMODS,SUBJECT=NP)
```

Esta regla, que mostramos de una forma muy simplificada, es la de sujeto. Toma un NP y un VP y los une, en el caso de que se cumplan determinadas condiciones. Una de las condiciones fundamentales que se deben cumplir es que exista concordancia en número y persona entre el sujeto y el verbo. Si esto es así, se crea un nuevo elemento.

Cuando SPG produce un análisis sintáctico de una frase, aparece en la pantalla un árbol. El árbol tiene la ventaja de que es más fácil de leer y cualquiera puede interpretarlo. Sin embargo, tiene menos información que los registros. Lo veremos enseguida.

Para el ejemplo siguiente obtenemos este árbol:

```
 Los chicos juegan en el jardin .

-------------------------------------------------------
DECL1  NP1      AJP1     ADJ1*   "Los"
                NOUN1*   "chicos"
       VERB1*   "juegan"
       PP1      PREP1    "en"
                AJP2     ADJ2*   "el"
                NOUN2*   "jardin"
       PUNC1    "."
-------------------------------------------------------
```

Si observamos la información que llevan los registros, comprobamos que es mucho más rica que la que proporciona el árbol.

```
SEGTYPE    'SENT'
SEGTYP2    'DECL'
STR        " los chicos juegan en el jardin ."
RULES      2110 2510 3000 5320
RULE       5320 SNTBEG1 VP1 PUNC1
COPYOF     VP1 "los chicos juegan en el jardin" 'JUGAR'
BASE       'JUGAR'
DICT       'juegan'
INDIC      PLUR PERS3 PRES INDICAT
PRMODS     NP1 "los chicos" 'CHICO'
HEAD       VERB1 "juegan" 'JUGAR'
PSMODS     PP1 "en el jardin" 'JARDIN'
PSMODS     PUNC1 "." '.'
SUBJECT    NP1 "los chicos" 'CHICO'
PARSENO    1
NODENAME   'DECL1'
```

En la parte izquierda aparecen los atributos y, a su derecha, el valor. Los valores pueden ser simples o complejos y muchos registros pueden ser atributos a su vez. Señalamos los atributos esenciales que aparecen: **PRMODS, HEAD, PSMODS, SEGTYPE y STR.**

Los otros atributos también muestran mucha información. Por ejemplo, **RULES** dice qué reglas se han utilizado en el análisis; **INDIC** muestra los rasgos de palabras o frases; **BASE** muestra la base del núcleo de un constituyente; **POS** informa sobre las posibles partes de la oración para una palabra.

Además, la información funcional, si está disponible, también aparece (en este caso **SUBJECT**).

Las reglas se aplican *bottom-up* y en paralelo. Esto significa que se aplicarán todas las reglas en las que haya condiciones que se cumplan, pudiendo llegar a producirse, para una única cadena, varios análisis, si no existen restricciones que lo impidan.

Lo más deseable es que aparezca un único análisis para cada frase (excepto cuando haya ambigüedad real, en cuyo caso será correcto que aparezcan varios análisis), pero si aparecen varios, existe una función (**P-METRIC**) que los maneja y establece un orden de prioridad para ellos. Otras veces, no se llega a producir ningún análisis (por ejemplo, cuando hay un error gramatical). Si esto ocurre (imaginemos un caso en el que haya falta de concordancia entre el sujeto y el verbo), las condiciones que afectan a la concordancia se relajan en un segundo intento de análisis y si, como consecuencia de esa relajación, conseguimos un análisis es que, efectivamente, estamos ante un error de falta de concordancia y es cuando se le muestra al usuario. Si, por el contrario, no obtenemos tampoco un análisis, se aplica otra función (**fitted**) que permite, al menos, que consigamos un análisis parcial. Como vemos, y hemos señalado anteriormente, los errores gramaticales se identifican en ese segundo paso, como una parte del análisis sintáctico.

Las reglas que manejan los problemas estilísticos se detectan mediante funciones que se escriben también en PLNLP. El cuerpo de la función tiene casi la misma forma que la parte derecha de las reglas de la gramática (donde se especifican las acciones o creación de nuevos elementos). Un ejemplo resumido sería:

```
STYLE (SEG*PTR,
    <SEGTYP2(TOP<SENTLIST(MEM)>).ISIN.&(XXXX XXNP), +FIT>,
    SERICOMA<SEG>,STPUNCI<SEG>)
```

A la izquierda, aparece el nombre de la función y, a continuación, se declaran los parámetros y se escribe el cuerpo. En este caso, estamos diciendo que si, después de aplicar las reglas de la gramática, no hemos conseguido un árbol completo del tipo SENT para una oración determinada, estamos ante un análisis parcial, (**fitted**). En la última línea de esta función se invoca a otras dos, SERICOMA y STPUNCI .

Conclusión

En la actualidad, contamos con un prototipo de nuestro sistema CORRECTOR que es capaz de analizar frases bastante complejas y de todo tipo (oraciones de relativo, subordinadas, con aposiciones, coordinadas, perífrasis...) y de detectar y proporcionar información para algunos errores gramaticales y problemas estilísticos, como hemos visto en este artículo. En un futuro próximo, a la vez que seguimos trabajando en el desarrollo de la gramática propiamente dicha, seguiremos aumentando el espectro de errores (modo verbal incorrecto, usos erróneos de preposiciones...), para intentar reflejar los errores más frecuentes en un texto escrito.

Referencias

[1] CASAJUANA, R., RODRÍGUEZ, C.: "Verificación ortográfica en castellano; la realización de un diccionario en ordenador", Español Actual, no. 44,1985.

[2] CASAJUANA, R., RODRÍGUEZ, C., SOPEÑA, L., VILLAR, C.: "Towards an Integrated Environment for Spanish Document Verifl cation and Composition", Proc. 3rd. Conference Association for Computational Linguistics (European Chapter), Copenhagen, April 1987, S2-55.

[3] RODRÍGUEZ, C., SOPEÑA, L., VILLAR, C., "A Computer Implementation of a Spanish Synonym Dictionary", **Literary and Linguistic Computing,** vol. 4, no. 4, pp. 265-270.

[4] RODRÍGUEZ, C., SOPEÑA, L., VALLADARES, C., VILLAR, C., "Clasificación morfológica del léxico castellano para un analizador en ordenador", VII Congreso de AESLA, Sevilla, abril 1989.

[5] RODRÍGUEZ, C., SOPEÑA, L., VALLADARES, C., VILLAR, C., "A Lexical Data Base of Spanish for Natural Language Applications" The Dynamic Text Conference, 16th ALLC Conference, Toronto, Canadá, junio 1989.

[6] ALCALÁ, A., RODRÍGUEZ, C., SOPEÑA, L., VILLAR, C., "Elaboración de una codificación sintáctica en ordenador de los verbos castellanos~, V Congreso de Lenguajes Naturales y Lenguajes Formales, Barcelona, septiembre 1989.

[7] RODRÍGUEZ, C., SOPEÑA, L., VILLAR, C., "Confección de un diccionario de sinónimos en ordenador. Teoría, metodología y resultados~, Linguística Española Actual, 1990.

[8] RODRÍGUEZ, C., "Introducción a SPG. Entorno y herramientas". Centro de Investigacion UAM-IBM, 1991. (En preparación)

[9] RODRÍGUEZ, C., "SPS (Spanish PLNLP Grammar): Elaboración de una gramática española". Centro de Investigación UAM-IBM, 1991. (En preparación)

[10] EL PAÍS. Libro de estilo. Ediciones EL PAÍS. Madrid, 1990.

[11] SECO, Manuel. Diccionario de dudas y dificultades de la lengua española. Espasa-Calpe, Madrid, 1986.

[12] MOLINER, M. Diccionario de uso del español. 2 vols. Gredos. Madrid, 1982.

[13] ALARCOS, Emilio. Estudios de Gramática Funcional del español. Gredos, Madrid, 1978.

[14] DUBOIS, Jean y otros. Diccionario de Lingüística. Alianza Editorial, Madrid, 1986.

[15] SECO, M. Gramática esencial del español. Introducción al estudio de la lengua. Espasa-Calpe, Madrid, 1989.

[16] HEIDORN, G. "Natural Language Inputs to a Simulation Programming System". Technical Report NPS-55HD72101A. 1972, Naval Postgraduate School. Monterey, CA.

[17] HEIDORN, G. The Programming Language for Natural Language Processing: PLNLP Language Reference Manual. NLP Systems Development Department, Bethesda, MD. DRAFT. IBM Internal Use Only, August, 1990.

[18] HEIDORN, G. "Augmented Phrase Structures Grammars~ in Nash-Webber and Schank, eds, **Theoretical Issues in Natural Language Processing** ACL, 1975.