

O conteúdo do presente relatório é de única responsabilidade do(s) autor(es).
(The contents of this report are the sole responsibility of the author(s).)

**Cálculo de la Estructura de un Texto en un
Sistema de Procesamiento de Lenguaje
Natural**

Horacio Saggion y Ariadne Carvalho

Relatório Técnico DCC-95-23

Dezembro de 1995

Cálculo de la Estructura de un Texto en un Sistema de Procesamiento de Lenguaje Natural*

Horacio Saggion y Ariadne Carvalho

Departamento de Ciência da Computação
Universidade Estadual de Campinas
Caixa Postal 6065
13081-970 Campinas-SP-Brazil
{saggion,ariadne}@dcc.unicamp.br

Abstract

Calcular la estructura lingüística de un texto es fundamental en cualquier sistema de procesamiento de lenguaje natural. Para eso, fenómenos mas allá de la barrera de la sentencia deben ser considerados. En particular, dos estudios deben ser abordados en el procesamiento de un texto: la cohesión textual y la coherencia textual. En este trabajo definimos una representación para la estructura de un resumen en lengua portuguesa considerado como un fenómeno lingüístico. La representación fue obtenida a partir del estudio de resúmenes reales y de la verificación de relaciones entre sentencias en el texto. Consideramos que este tipo de representación puede ser utilizada para otros textos, ya que es definida en función de relaciones de cohesión y coherencia. El principal aspecto considerado en este trabajo para la construcción de la estructura textual es la cohesión a través de la resolución de anáforas definidas tanto pronominales como frases nominales definidas. Presentamos ejemplos de texto reales y su tratamiento en el marco teórico propuesto.

Palabras clave: Procesamiento de Lenguaje Natural. Resúmenes. Estructura Textual. Cohesión Textual. Anáfora Definida.

Keywords: Natural Language Processing. Summaries. Textual Structure. Textual Cohesion. Definite Anaphora.

1 Introducción

Una gramática intenta capturar los conocimientos lingüísticos que un hablante tiene sobre su lengua y la lingüística tiene como objetivo el establecimiento de un conjunto de principios y reglas por medio de los cuales sea posible distinguir entre expresiones bien y mal formadas. Las aplicaciones en Lingüística Computacional se han concentrado en los últimos años en

*Este trabajo está siendo financiado por la fundación FAPESP

el texto como fenómeno lingüístico diferente de la sentencia. Sin embargo no existe por el momento un formalismo gramatical que permita distinguir entre textos bien y mal formados [Hovy 1993]. Para procesar adecuadamente un texto deben ser considerados dos aspectos fundamentales direccionados por la lingüística textual como bases para el cálculo de la estructura de un texto [Lopes Fávero 1991]: (i) la *cohesión* textual; (ii) y la *coherencia* textual. En este trabajo direccionamos el problema de la estructura textual de resúmenes de artículos técnicos escritos en lengua portuguesa. Definimos la estructura textual de un resumen y presentamos uno de los procesos que hacen posible el cálculo de esa estructura. El sistema completo propuesto aparece representado en la Figura 1. El texto es procesado con el objetivo de construir una representación estructural.

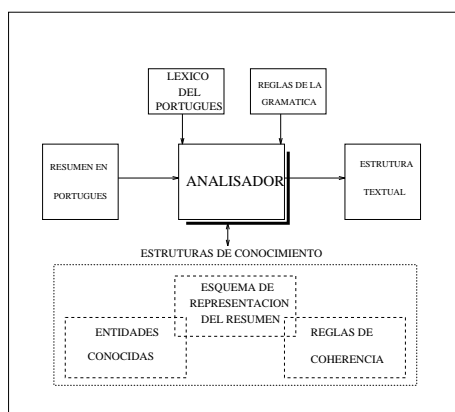


Figure 1: Esquema general del analizador textual

El principal aspecto considerado en este trabajo es la cohesión textual como un factor determinante en el cálculo de la estructura de un texto [Halliday y Hasan 1976].

El artículo se organiza de la siguiente manera: en la Sección 2 introducimos los conceptos relativos a resúmenes de artículos técnicos; en la Sección 3 presentamos la estructura textual de los resúmenes; en la Sección 4 introducimos los conceptos de cohesión textual. En la Sección 5 presentamos los procesos computacionales para el cálculo de la cohesión textual y finalmente en la Sección 6 son presentadas las conclusiones de este trabajo.

2 Resúmenes

Un resumen es la primera parte de un reporte técnico; según la ciencia de la documentación tratase de un indicador del contenido del reporte [Weil et al. 1963a, Cleveland y Cleveland 1983]. Según la lingüística [Gopnik 1972, Hutchins 1985], a pesar de ser restricto a un único párrafo, tratase de un fenómeno textual y presenta todas las relaciones interesantes de textos mayores, en particular relaciones de cohesión y de coherencia.

La ciencia de la documentación distingue entre dos tipos básicos de resúmenes [Jordan 1991, Cleveland y Cleveland 1983]: (i) *indicativo o descriptivo*, considerado como una tabla de contenidos del reporte técnico pero escrito en lenguaje natural; indica el material que será

encontrado en el documento original y jamás podrá transformarse en un sustituto de este último; e (ii) *informativo*, considerado un verdadero texto ya que presenta datos e informaciones del propio reporte, contiene las ideas y los hechos principales del documento y muchas veces puede ser utilizado como sustituto del documento original. En la práctica, un resumen será un híbrido entre los dos tipos considerados.

Resúmenes de los campos de conocimiento mas variados son escritos de manera similar; existen además publicaciones que establecen algunas guías sobre como este tipo de texto debe ser escrito y sobre el tipo de información a ser incluida [Weissberg y Buker 1990, Weil et al. 1963b].

Un resumen incluirá en general los siguientes ítems de información: (i) *Contexto*, donde información preliminar es incluida y el problema estudiado es definido; (ii) *Objetivos y alcance* del estudio realizado; (iii) *Metodología* empleada en la investigación, incluyendo materiales, equipamientos y procedimientos utilizados; (iv) *Resultados* obtenidos en la investigación; y (v) *Conclusiones* alcanzadas en el trabajo, incluyendo recomendaciones y sugerencias.

Además del tipo de información que aparecerá en un resumen, también pueden ser verificadas algunas normas de utilización de tiempos verbales que corresponden con una situación pragmática; el tiempo pasado es utilizado para describir informaciones relacionadas con el trabajo científico, dado que este fue realizado con anterioridad a la redacción del reporte y el tiempo presente es utilizado para describir las conclusiones y los objetivos del trabajo ya que son informaciones asociadas directamente con el documento.

Nuestro trabajo se basa en el análisis de un “corpus” de resúmenes reales publicados por la “Revista de Ensino de Engenharia”. Escogimos estudiar textos de esa publicación, pues los resúmenes publicados deben seguir las normas de la “Associação Brasileira de Normas Técnicas” [ABNT 1987]. El estudio de textos reales permite la observación de fenómenos lingüísticos generales y particulares así como la comprobación de los aspectos teóricos estudiados. En la Figura 2, presentamos un resumen del corpus estudiado [Milanez 1987].

En este resumen podemos verificar la inclusión de los siguientes tipos de información: las

(1) Os programas de avaliação de docentes pelos alunos vêm sendo utilizados nos Estados Unidos desde a década de quarenta. (2) Este tipo de programa não é muito comum nas universidades brasileiras. (3) No Departamento de Engenharia Mecânica da UNICAMP foram feitas algumas tentativas de implantação destes programas. (4) O presente trabalho descreve a experiência da aplicação destes programas no Departamento de Engenharia Mecânica da UNICAMP.

Figure 2: Un resumen

sentencias (1) y (2) presentan información de tipo contextual. La sentencia (3) presenta información sobre el trabajo científico y la sentencia (4) presenta información sobre el objetivo del reporte técnico.

3 La estructura textual

La estructura de un texto ha sido estudiada desde diversas perspectivas. [McKeown 1985] estudió la estructura de textos cortos producidos en respuesta a preguntas sobre la organización de un banco de datos. Ella notó que el tipo de texto producido en dicha situación sigue padrones muy estereotípicos. También notó que en textos científicos los autores utilizan conocimientos sobre el tipo de material que debe aparecer en la introducción y los datos a ser incluidos en la conclusión. [Gopnik 1972] estudió la estructura de textos científicos interesada en reconocer las características sintácticas de las sentencias en el contexto textual. [Grosz y Sidner 1986] propusieron la estructura de un discurso de modo tal que el modelo actúa para restringir los posibles antecedentes para un pronombre.

Algunos autores reconocen que la estructura de un texto está dada por las relaciones que vinculan sentencias adyacentes en la superficie lingüística. [Mann y Thompson 1987] estudiaron textos reales y definieron una teoría que permite describir la estructura de un texto en términos de relaciones entre las sentencias. La mayor parte de las relaciones tiene una palabra clave para facilitar su identificación. La teoría, por ellos propuesta, fue utilizada en el campo de generación de textos por computador [Hovy 1993, Moore y Paris 1994].

3.1 La estructura de un resumen

Definimos la estructura de un resumen como formada por tres partes fundamentales:

- (i) una partición del texto en grupos de sentencias adyacentes en la superficie lingüística. Llamaremos a cada grupo *segmento*. El nombre de una *Categoría de Información* es asignado al segmento.
- (ii) un conjunto de *proposiciones* y *entidades* producidas a partir del análisis sintáctico-semántico de las sentencias.
- (iii) un conjunto de *relaciones proposicionales* que vinculan las proposiciones.

3.1.1 Categorías de Información

Las *Categorías de Información (CI)* por nosotros utilizadas para identificar los segmentos son:

$$CI \equiv \{ \textit{Contexto}, \textit{Objetivo}, \textit{Método}, \textit{Experimento}, \textit{Resultado}, \textit{Recomendación}, \textit{Sugestión}, \textit{Conclusión} \}$$

El nombre de la categoría indica el tipo de material contenido en el segmento dado que cada segmento en un texto debe servir a un propósito de información [Grimes 1975]. Para la determinación de los segmentos utilizamos dos tipos de información: (i) la categoría lingüística *tiempo* que es considerada el principal indicador de cambio de segmento y (ii) *información semántica* asignada a los ítems de la lengua que fueron identificados como indicadores de segmentos.

3.1.2 Entidades

Cada sintagma nominal con m modificadores será representado por una expresión de la forma (1).

$$p(i, e_1, \dots, e_m, d, g, n) \quad (1)$$

Donde p es un predicado nominal asociado a la cabeza nominal del sintagma; i es un índice que identifica el referente del sintagma nominal; e_k es una expresión que representa el k -ésimo complemento del sintagma nominal; y d, g y n son respectivamente expresiones asociadas a la información sobre determinación, género y número. Las expresiones de los complementos pueden ser *modificador* = *constante_{nominal}*, *calificador* = *constante_{adjetiva}*, *especificador* = *índice* dependiendo del complemento. La asociación de cada complemento nominal a un elemento del lenguaje de representación es definido en el léxico de la lengua. La representación es obtenida por un algoritmo de análisis sintáctico-semántico para las frases nominales del portugués.

3.1.3 Proposiciones

Cada verbo con n argumentos será representado por una expresión del tipo (2).

$$p(i, papel_1 = e_1, \dots, papel_n = e_n, t, a, m) \quad (2)$$

Donde p es un predicado asociado al verbo; $papel_k$ es el papel temático que el k -ésimo complemento desempeña en la estructura de la frase; e_k es el índice del k -ésimo complemento del verbo y t, a y m representan respectivamente las informaciones sobre tiempo, aspecto y modo asociados al verbo.

Obtenemos las proposiciones utilizando un algoritmo de análisis sintáctico guiado por información sobre los posibles complementos de un verbo (información sobre el marco de sub-categorización del verbo).

3.1.4 Relaciones Proposicionales

Estamos trabajando con el siguiente conjunto de *Relaciones Proposicionales (RP)*;

$$RP \equiv \{ \textit{Elaboración}, \textit{Paralelo}, \textit{Secuencia}, \textit{CausaConsecuencia}, \textit{Contraste}, \textit{SecuenciaTemporal}, \textit{Ocasión} \}$$

Las relaciones *Elaboración* y *Paralelo* son identificadas por los vínculos cohesivos entre entidades en proposiciones adyacentes. Las restantes relaciones son identificadas según las construcciones sintácticas de la lengua portuguesa, utilizadas en las sentencias.

Las relaciones asociadas a construcciones sintácticas son deducidas en la fase de análisis sintáctica de la sentencia y dependen de la forma de la frase.

La representación completa del resumen de la Figura 2 puede ser observada en la Figura 3.

Para calcular la relación de *Elaboración* que vincula las oraciones (1) y (2) es imprescindible verificar las relaciones cohesivas que existen entre estas sentencias. Por eso direccionamos en este trabajo el análisis de la cohesión textual como paso fundamental en el cálculo de la estructura de un texto.

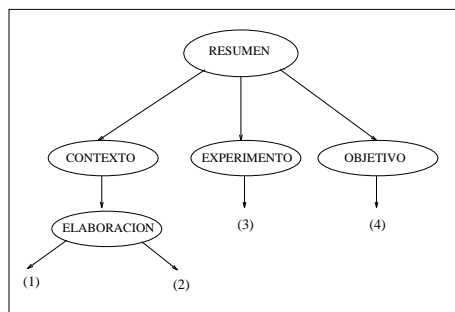


Figure 3: Estructura textual del resumen

4 La cohesión textual

Existen mecanismos que los escritores utilizan para señalar relaciones de sentido entre partes de enunciados. Es por medio de esos mecanismos que la estructura textual es construida. [Koch 1989] distingue para el portugués dos tipos de cohesión: (i) la cohesión referencial; y (ii) la cohesión secuencial. En este trabajo estaremos concentrados en la cohesión referencial como mecanismo de construcción de la estructura de un texto.

Cuando escribimos utilizamos expresiones para referirnos a entidades y situaciones del mundo real, el problema de determinar el referente de una expresión lingüística es el clásico problema de la referencia semántica (exófora). En portugués, las entidades con poder de referencia son los sintagmas nominales que designan entidades y objetos de la situación comunicativa y las oraciones que designan propiedades y eventos [Mateus et al. 1983].

Existen situaciones en las cuales los escritores utilizan *formas abreviadas* para referir a una entidad previamente introducida en el contexto discursivo [Hirst 1981]. La forma abreviada es llamada ítem *remisivo* y el elemento al cual este refiere es llamado *antecedente*. El proceso por el cual se determina el antecedente de un ítem remisivo se llama *resolución*. Cuando se establece una relación entre ambos se dice que esos elementos (remisivo y antecedente) son *co-referentes*. Dos tipos de co-referencia son identificados: (i) cuando el ítem remisivo aparece antes que su antecedente en la secuencia lingüística estamos en un caso de *catáfora*; (ii) cuando el ítem remisivo aparece después de su antecedente en la secuencia lingüística estamos en un caso de *anáfora*. En este trabajo consideraremos únicamente casos de co-referencia anafórica. En estos casos el ítem remisivo es también llamado *anáfora*.

El caso más evidente de anáfora es la de los pronombres personales de tercera persona ¹, los pronombres reflexivos², los pronombres objeto³ y los pronombres demostrativos⁴. Estas formas abreviadas no tienen ninguna instrucción semántica que permita la determinación de su antecedente; apenas instrucciones de conexión son especificadas en el léxico de la lengua. No obstante, varias son las restricciones estructurales impuestas por la teoría gramatical [Raposo 1992, Lobato 1986] para la elección o exclusión de un antecedente para un

¹ ele(s)/ela(s)

² se/si próprio/si própria

³ o(s)/a(s)

⁴ este(s)/esta(s),esse(s)/essa(s),aquele(s)/aquela(s)

pronombre. Consideremos por ejemplo la sentencia (a) de la Figura 4 y su representación en términos de categorías sintácticas.

(a) a engenharia depende das ciências e de [técnicas n[elas] _{NP₂} baseadas] _{NP₁}
(b) a engenharia não prescinde d[a criatividade de quem [a] _{NP₂} usa] _{NP₁}

Figure 4: Filtro *i* sobre *i* para un pronombre

El pronombre personal *elas* (NP_2) no puede tener como antecedente el sintagma nominal completo cuya cabeza es *técnicas* (NP_1), a pesar del pronombre y el sintagma compartir concordancia en género y número. En este caso se impone la restricción gramatical: *Filtro *i* sobre *i**, por medio de la cual una expresión nominal A no puede ser co-referente con otra expresión nominal B de la cual A forma parte. Otro caso de esta restricción puede ser observado en la sentencia (b) de la Figura 4, donde el pronombre objeto a (NP_2) no puede ser co-referente con el sintagma cuya cabeza es *criatividade* (NP_1), también por la misma restricción estructural. Las restricciones impuestas por la gramática hacen posible la exclusión de pares de expresiones nominales a nivel sentencial y también hacen posible la determinación de posibles referentes para ítems reflexivos. En el nivel textual la resolución de los pronombres se realiza considerando las entidades previamente introducidas en el texto y que concuerdan en género y número con el pronombre. Sin embargo, existen autores que consideran la propia estructura del discurso como determinante en la elección de un antecedente para un pronombre [Grosz y Sidner 1986, Brennan et al. 1987].

El segundo tipo de elemento anafórico es la frase nominal definida (FND). Una FND es aquel sintagma nominal al que han sido aplicadas operaciones de determinación por medio de los determinantes del portugués: $\{o, este, esse, aquele\}$ y sus respectivos femeninos y plurales. No todas FND son consideradas como elementos anafóricos. Categorizamos las FND según el siguiente criterio: (i) FND que refieren a la situación comunicativa y que son consideradas como únicas; (ii) FND con operaciones de especificación que refieren al contexto o están siendo introducidas por primera vez en el texto; y (iii) FND sin operaciones de complementación, pero consideradas incompletas. Las frases clasificadas como (i) y (ii) deben encontrar su referente en la base de conocimientos del sistema o debe crearse un referente en caso de ser introducidas por primera vez. Las frases clasificadas como (iii) deben encontrar un elemento co-referente en el propio texto, el cual debe anteceder al ítem remitivo.

En este trabajo estamos considerando referencias a elementos previamente introducidos considerando únicamente la cabeza nominal de la FND, otros casos de co-referencia no serán aquí tratados.

Consideremos el segmento de texto de la Figura 5 [Gomide y Fernández 1985]. En la sentencia (1) la expresión “a disciplina Similitude em Engenharia” introduce una entidad en el entorno discursivo. En la sentencia (2) esta entidad es anafóricamente retomada por medio de la expresión abreviada “esta disciplina”. Para resolver esta referencia abreviada es utilizado el predicado nominal *disciplina* asociado con la cabeza nominal de la expresión

(1) *O presente trabalho tem como meta divulgar a disciplina Similitude em Engenharia ministrada no curso de Engenharia Mecânica da UFU.* (2) *Esta disciplina visa proporcionar aos alunos fundamentos básicos sobre a teoria de modelos.*

Figure 5: Referencia anafórica en texto

abreviada.

Consideremos la expresión nominal “estes programas” en la sentencia (3) del resumen de la Figura 2. Esa frase nominal definida es una referencia abreviada que tiene como antecedente la expresión “os programas de avaliação de docentes pelos alunos” de la sentencia (1). La resolución de la referencia abreviada es realizada utilizando el predicador *programa* asociado a la cabeza nominal de la expresión abreviada.

5 Resolución de la cohesión referencial

Para resolver las referencias pronominales en el texto, nos hemos basado en una estructura de pila que mantiene las entidades introducidas en el discurso en el orden en que aparecieron: la más reciente será la más accesible. La resolución de las frases nominales definidas está basada en una base de conocimientos que contiene referentes, reglas de vinculación conceptual y referentes activos.

5.1 Resolución de los pronombres

En este trabajo estamos considerando únicamente los pronombres personales y los pronombres objeto de tercera persona, los pronombres reflexivos y recíprocos de la lengua portuguesa no serán tratados en este artículo. Un pronombre es considerado una referencia anafórica, por lo tanto su antecedente deberá ser una entidad introducida previamente en el texto. Utilizamos una pila AP para el almacenamiento de las entidades introducidas en el texto y que sirven de antecedentes para un pronombre. Cada elemento de la pila es una 3-upla de *índice*, *género* y *número*, y las operaciones sobre la pila son las usuales: *apilar*, *desapilar* y *tope*. Cuando una nueva entidad es introducida en el texto, y portanto producida una representación de la forma $p(i_e, \dots, d, g, n)$, la pila AP es actualizada según (3).

$$AP \leftarrow \text{apilar}(AP, \langle i_e, g, n \rangle) \quad (3)$$

Cuando un pronombre es encontrado en el texto se produce una expresión del tipo (4).

$$\text{anafora}(i_a, g_a, n_a) \quad (4)$$

El antecedente del pronombre se buscará utilizando el esquema presentado en la Figura 6. Podemos observar dos procesos principales: (i) cálculo de las entidades no-coreferenciales

aplicando restricciones estructurales; y (ii) búsqueda del antecedente.

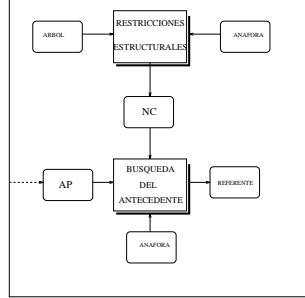


Figure 6: Resolución de pronombres

En el proceso (i) es obtenido el conjunto de índices que no pueden ser antecedentes para el pronombre; este paso es realizado considerando restricciones en el nivel sentencial aplicadas al árbol de análisis sintáctico. Para cada anáfora i_a es calculado el conjunto $NC(i_a)$ de pares no co-referenciales según la especificación (5).

$$NC(i_a) \equiv \{i_{a_1}, \dots, i_{a_k}\} \quad (5)$$

tal que i_{a_j} ($1 \leq j \leq k$) es no co-referente con i_a

El paso (ii) de búsqueda del antecedente para el pronombre se basa en la especificación (6)⁵. La pila AP es examinada hasta obtener un posible antecedente para el pronombre considerando trazos sintácticos.

$$X \in AP / \Pi_1(X) \notin NC(i_a) \wedge \Pi_2(X) = g_a \wedge \Pi_3(X) = n_a \quad (6)$$

Consideremos como ejemplo la resolución del pronombre *elas* de la sentencia (a) de la Figura 4. El estado de la pila en ese momento puede ser observado en la Figura 7. Asumimos que los referentes de “a engenharia”, “as ciencias” y “técnicas” son respectivamente i_0 , i_1 e i_2 .

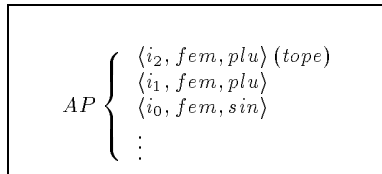


Figure 7: Estado de la pila (a)

⁵La operación Π_k ($1 \leq k \leq 3$) proyecta el k -ésimo componente de la upla X

La representación del pronombre en el lenguaje de representación es (7) y el conjunto de pares no co-referenciales es (8).

$$anafora(i_a, fem, plu) \tag{7}$$

$$NC(i_a) = \{i_2\} \tag{8}$$

El proceso de búsqueda del antecedente para el pronombre finaliza con la determinación de i_1 como co-referente con i_a .

Consideremos ahora la resolución del pronombre objeto a de la sentencia (b) de la Figura 4. El estado de la pila es mostrado en la Figura 8. Asumimos que los referentes de “a engenharia” y “a criatividade” son respectivamente i_0 e i_3 .

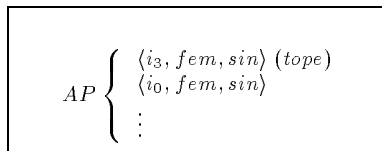


Figure 8: Estado de la pila (b)

La representación del pronombre a en el lenguaje de representación es (9) y el conjunto de pares no co-referenciales es (10).

$$anafora(i_b, fem, sin) \tag{9}$$

$$NC(i_b) = \{i_3\} \tag{10}$$

El proceso de búsqueda del antecedente para el pronombre finaliza con la determinación de i_0 como co-referente con i_b .

5.2 Resolución de FND

Para la resolución de las FND utilizamos una base de conocimientos que contiene referentes para las FND consideradas completas y reglas para la vinculación conceptual de las entidades en el contexto discursivo. El referente de una FND de tipo (i) es buscado en la base de conocimientos, cuando el referente de la expresión definida es encontrado este es marcado como activo. El referente de una FND de tipo (ii) es buscado utilizando reglas que establecen relaciones entre entidades introducidas en el discurso, y por lo tanto activas, y entidades a ellas conceptualmente vinculadas. Una FND de tipo (iii) es considerada una referencia abreviada, el antecedente de esta expresión será buscado considerando el predicado asociado a la cabeza de la FND.

Así, por ejemplo considerando el segmento de texto de la Figura 5, la expresión “a disciplina Similitude em Engenharia” es considerada una FND de tipo (i), luego, su referente

será buscado en la base de conocimiento del sistema utilizando la expresión de búsqueda (11).

$$disciplina(X, nome = similitude_em_engenharia) \quad (11)$$

Supongamos que el índice de ese referente es i_0 , este índice será marcado como activo. La expresión “esta disciplina” es considerada una FND de tipo (iii), luego, su antecedente deberá ser buscado considerando el predicado *disciplina*. La búsqueda en la base de conocimiento se realiza utilizando una expresión del tipo (12).

$$disciplina(X) \wedge activa(X) \quad (12)$$

La base de conocimientos responde a esa búsqueda con $X = i_0$. Considerando que el referente i_0 es introducido en la sentencia (1) del segmento y que este referente es retomado en la sentencia (2), podemos verificar la relación de *Elaboración* en el segmento.

Consideremos ahora la expresión definida “o plano de ensino” que es considerada una FND de tipo (ii). El referente de esta expresión será buscado en la base de conocimiento utilizando la regla (13) y la expresión de búsqueda (14).

$$plano(p(X), modificador = ensino) \Leftarrow disciplina(X) \wedge activa(X) \quad (13)$$

$$plano(X, modificador = ensino) \quad (14)$$

La solución encontrada es $X = p(i_0)$. Este referente debe ser marcado como activo en la base de conocimiento.

Como último ejemplo consideremos la resolución de las frases nominales definidas del resumen de la Figura 2. En este texto observamos tres frases nominales definidas relacionadas: (a) “os programas de avaliação de docentes pelos alunos” en la sentencia (1); (b) “este tipo de programa” en la sentencia (2); y (c) “estes programas” en las sentencias (3) y (4). Para la expresión (a) es encontrado en la base de conocimientos un referente i_p ; este referente será marcado como activo para una posible referencia posterior y también para activar las reglas que vinculan este referente con otras entidades. La expresión (b), dadas sus características de complementación y la información tomada del léxico, es buscada utilizando la regla (15) y la expresión de búsqueda (16).

$$tipo(X, modificador = programa) \Leftarrow programa(X) \wedge activa(X) \quad (15)$$

$$tipo(X, modificador = programa) \quad (16)$$

La base de conocimientos responde con $X = i_p$. Finalmente la expresión (c) es considerada una referencia abreviada por lo tanto será buscada en la base de conocimiento utilizando la expresión (17).

$$programa(X) \wedge activa(X) \quad (17)$$

La base de conocimientos responderá con $X = i_p$.

6 Conclusiones

En este trabajo hemos abordado la cohesión textual como un paso fundamental en el cálculo de la estructura de un texto. El tipo de texto considerado en este trabajo está restringido a resúmenes de artículos técnicos que, a pesar de ser restringidos a un único párrafo, son considerados por la lingüística como un fenómeno textual. Hemos definimos un formalismo para la representación de este tipo de texto y hemos presentado el proceso que hace posible la verificación de relaciones cohesivas. El formalismo de representación puede ser extendido para abarcar otros tipos de texto visto que las relaciones establecidas entre sentencias son de tipo lógico-semántico y de tipo cohesivo. En este trabajo hemos direccionado un tipo de referencia anafórica conocido como anáfora definida y hemos presentado los mecanismos computacionales que posibilitan la resolución de algunos pronombres del portugués y de algunas frases nominales definidas. No hemos considerado el tratamiento de otros tipos de pronombres como los pronombres substantivos demostrativos. En lo que concierne a la resolución de las frases definidas anafóricas consideramos únicamente el predicado asociado a la cabeza nominal de la frase definida, si bien en el “corpus” estudiado este es el mecanismo mas frecuente, no es el único, por lo tanto investigación adicional es necesaria para tratar este y otros tipos de referencia que aparecen en la literatura.

References

- [ABNT 1987] ABNT-Associação Brasileira de Normas Técnicas. *Resumos*. 1987.
- [Brennan et al. 1987] Brennan, S., Friedman, M.W. y Pollard, C.J. *A Centering Approach to Pronouns*. 25th Annual Meeting of the ACL, 6-9 July 1987.
- [Cleveland y Cleveland 1983] Cleveland, D.B. y Cleveland A.D. *Introduction to Indexing and Abstracting*. Libraries Unlimited, Inc., 1983.
- [Gomide y Fernández 1985] Gomide, H.A. y Fernández, E.F. *Curso de Similitude em Engenharia*. Revista de Ensino de Engenharia, 4(2):125-132, São Paulo, 2^o Sem. 1985.
- [Gopnik 1972] Gopnik, M. *Linguistic Structures in Scientific Texts*. Mouton, 1972.
- [Grimes 1975] Grimes, J. *The Thread of Discourse*. Mouton and Company, The Hague, Netherlands, 1975.
- [Grosz y Sidner 1986] Grosz, B.J. y Sidner, C.L. *Attention, Intentions and the Structure of Discourse*. Computational Linguistics, Vol. 12, Num. 3, July-September 1986.
- [Halliday y Hasan 1976] Halliday, M.A. y Hasan, R. *Cohesion in English*. London, Longman Press, 1976.
- [Hirst 1981] Hirst, G. *Anaphora in Natural Language Understanding: A Survey*. Lecture Notes in Computer Science 119. Springer-Verlag, 1981.

- [Hovy 1993] Hovy, E. *Automated Discourse Generation using Discourse Structure Relation*. Artificial Intelligence, Vol. 63, Num. 1-2, October 1993.
- [Hutchins 1985] Hutchins, W.J. *Information Retrieval and Text Analysis*. In New Approaches to the Analysis of Mass, Media, Discourse and Communication. T.A. van Dijk (Ed.), Gruyter, Berlin, 1985.
- [Jordan 1991] Jordan, M.P. *The Linguistic Genre of Abstracts*. In A. Della Volpe (ed.), The Seventeenth LACUS Forum. Linguistics Association of Canada and the United States, 1991.
- [Koch 1989] Koch, I.V. *A Coesão Textual*. Editora Contexto, 1989.
- [Lobato 1986] Lobato, M.L.P. *Sintaxe Gerativa do Português. Da Teoria Padrão à Teoria da Regência e Ligação*. Vigília, 1986.
- [Lopes Fávero 1991] Lopes Fávero, L. *Coesão e Coerência Textuais*. Editora Ática, 1991.
- [Mann y Thompson 1987] Mann, W.C. y Thompson S.A. *Rhetorical Structure Theory: A Theory of Text Organization*. ISI Reprint Series, ISI/RS-87-190, June 1987.
- [McKeown 1985] McKeown, K.R. *Discourse Strategies for Generating Natural-Language Text*. Artificial Intelligence, Vol. 27, 1985.
- [Milanez 1987] Milanez, L.F. *Considerações sobre Programas de Avaliação de Docentes pelos Alunos*. Revista de Ensino de Engenharia, 6(1):47-52, São Paulo, 1^o Sem. 1987.
- [Mateus et al. 1983] Mateus, M.H.M., Brito, A.M., Duarte, I.S. y Faria, I.H. *Gramática da Língua Portuguesa*. Livraria Almedina, Coimbra, 1983.
- [Moore y Paris 1994] Moore, J.D. y Paris, C.L. *Planning Text for Advisory Dialogues: Capturing Intentional and Rhetorical Information*. Computational Linguistics, Vol. 19, Num. 4, 1994.
- [Raposo 1992] Raposo, E.P. *Teoria da Gramática. A Faculdade da Linguagem*. Ed. Caminho, Lisboa, 1992.
- [Weil et al. 1963a] Weil, B.H., Zarembler, I. y Owen, H. *Technical-Abstracting Fundamentals. Introduction*. Journal of Chemical Documentation, Num 3, 1963.
- [Weil et al. 1963b] Weil, B.H., Zarembler, I. y Owen, H. *Technical-Abstracting Fundamentals. Writing Principles and Practices*. Journal of Chemical Documentation, Num 3, 1963.
- [Weissberg y Buker 1990] Weissberg, R. y S. Buker. *Writing UP Research*. Prentice-Hall, Inc., 1990.

Relatórios Técnicos – 1992

- 92-01 **Applications of Finite Automata Representing Large Vocabularies,**
C. L. Lucchesi, T. Kowaltowski
- 92-02 **Point Set Pattern Matching in d -Dimensions,** *P. J. de Rezende, D. T. Lee*
- 92-03 **On the Irrelevance of Edge Orientations on the Acyclic Directed Two Disjoint Paths Problem,** *C. L. Lucchesi, M. C. M. T. Giglio*
- 92-04 **A Note on Primitives for the Manipulation of General Subdivisions and the Computation of Voronoi Diagrams,** *W. Jacometti*
- 92-05 **An (l, u) -Transversal Theorem for Bipartite Graphs,** *C. L. Lucchesi, D. H. Younger*
- 92-06 **Implementing Integrity Control in Active Databases,** *C. B. Medeiros, M. J. Andrade*
- 92-07 **New Experimental Results For Bipartite Matching,** *J. C. Setubal*
- 92-08 **Maintaining Integrity Constraints across Versions in a Database,**
C. B. Medeiros, G. Jomier, W. Cellary
- 92-09 **On Clique-Complete Graphs,** *C. L. Lucchesi, C. P. Mello, J. L. Szwarcfiter*
- 92-10 **Examples of Informal but Rigorous Correctness Proofs for Tree Traversing Algorithms,** *T. Kowaltowski*
- 92-11 **Debugging Aids for Statechart-Based Systems,** *V. G. S. Elias, H. Liesenberg*
- 92-12 **Browsing and Querying in Object-Oriented Databases,** *J. L. de Oliveira, R. de O. Anido*

Relatórios Técnicos – 1993

- 93-01 **Transforming Statecharts into Reactive Systems**, *Antonio G. Figueiredo Filho, Hans K. E. Liesenberg*
- 93-02 **The Hierarchical Ring Protocol: An Efficient Scheme for Reading Replicated Data**, *Nabor das C. Mendonça, Ricardo de O. Anido*
- 93-03 **Matching Algorithms for Bipartite Graphs**, *Herbert A. Baier Saip, Cláudio L. Lucchesi*
- 93-04 **A lexBFS Algorithm for Proper Interval Graph Recognition**, *Celina M. H. de Figueiredo, João Meidanis, Célia P. de Mello*
- 93-05 **Sistema Gerenciador de Processamento Cooperativo**, *Ivonne. M. Carrazana, Nelson. C. Machado, Célio. C. Guimarães*
- 93-06 **Implementação de um Banco de Dados Relacional Dotado de uma Interface Cooperativa**, *Nascif A. Abousalh Neto, Ariadne M. B. R. Carvalho*
- 93-07 **Estadogramas no Desenvolvimento de Interfaces**, *Fábio N. de Lucena, Hans K. E. Liesenberg*
- 93-08 **Introspection and Projection in Reasoning about Other Agents**, *Jacques Wainer*
- 93-09 **Codificação de Seqüências de Imagens com Quantização Vetorial**, *Carlos Antonio Reinaldo Costa, Paulo Lício de Geus*
- 93-10 **Minimização do Consumo de Energia em um Sistema para Aquisição de Dados Controlado por Microcomputador**, *Paulo Cesar Centoducatte, Nelson Castro Machado*
- 93-11 **An Implementation Structure for RM-OSI/ISO Transaction Processing Application Contexts**, *Flávio Morais de Assis Silva, Edmundo Roberto Mauro Madeira*
- 93-12 **Boole's conditions of possible experience and reasoning under uncertainty**, *Pierre Hansen, Brigitte Jaumard, Marcus Poggi de Aragão*
- 93-13 **Modelling Geographic Information Systems using an Object Oriented Framework**, *Fatima Pires, Claudia Bauzer Medeiros, Ardemiris Barros Silva*
- 93-14 **Managing Time in Object-Oriented Databases**, *Lincoln M. Oliveira, Claudia Bauzer Medeiros*
- 93-15 **Using Extended Hierarchical Quorum Consensus to Control Replicated Data: from Traditional Voting to Logical Structures**, *Nabor das Chagas Mendonça, Ricardo de Oliveira Anido*

- 93-16 ***ℒℒ – An Object Oriented Library Language Reference Manual***, *Tomasz Kowaltowski, Evandro Bacarin*
- 93-17 ***Metodologias para Conversão de Esquemas em Sistemas de Bancos de Dados Heterogêneos***, *Ronaldo Lopes de Oliveira, Geovane Cayres Magalhães*
- 93-18 ***Rule Application in GIS – a Case Study***, *Claudia Bauzer Medeiros, Geovane Cayres Magalhães*
- 93-19 ***Modelamento, Simulação e Síntese com VHDL***, *Carlos Geraldo Krüger e Mário Lúcio Côrtes*
- 93-20 ***Reflections on Using Statecharts to Capture Human-Computer Interface Behaviour***, *Fábio Nogueira de Lucena e Hans Liesenberg*
- 93-21 ***Applications of Finite Automata in Debugging Natural Language Vocabularies***, *Tomasz Kowaltowski, Cláudio Leonardo Lucchesi e Jorge Stolfi*
- 93-22 ***Minimization of Binary Automata***, *Tomasz Kowaltowski, Cláudio Leonardo Lucchesi e Jorge Stolfi*
- 93-23 ***Rethinking the DNA Fragment Assembly Problem***, *João Meidanis*
- 93-24 ***EGOLib — Uma Biblioteca Orientada a Objetos Gráficos***, *Eduardo Aguiar Patrocínio, Pedro Jussieu de Rezende*
- 93-25 ***Compreensão de Algoritmos através de Ambientes Dedicados a Animação***, *Rackel Valadares Amorim, Pedro Jussieu de Rezende*
- 93-26 ***GeoLab: An Environment for Development of Algorithms in Computational Geometry***, *Pedro Jussieu de Rezende, Welson R. Jacometti*
- 93-27 ***A Unified Characterization of Chordal, Interval, Indifference and Other Classes of Graphs***, *João Meidanis*
- 93-28 ***Programming Dialogue Control of User Interfaces Using Statecharts***, *Fábio Nogueira de Lucena e Hans Liesenberg*
- 93-29 ***EGOLib – Manual de Referência***, *Eduardo Aguiar Patrocínio e Pedro Jussieu de Rezende*

Relatórios Técnicos – 1994

- 94-01 **A Statechart Engine to Support Implementations of Complex Behaviour**, *Fábio Nogueira de Lucena, Hans K. E. Liesenberg*
- 94-02 **Incorporação do Tempo em um SGBD Orientado a Objetos**, *Ângelo Roncalli Alencar Brayner, Cláudia Bauzer Medeiros*
- 94-03 **O Algoritmo KMP através de Autômatos**, *Marcus Vinícius A. Andrade e Cláudio L. Lucchesi*
- 94-04 **On Edge-Colouring Indifference Graphs**, *Celina M. H. de Figueiredo, João Meidanis, Célia Picinin de Mello*
- 94-05 **Using Versions in GIS**, *Claudia Bauzer Medeiros and Geneviève Jomier*
- 94-06 **Times Assíncronos: Uma Nova Técnica para o Flow Shop Problem**, *Hélvio Pereira Peixoto e Pedro Sérgio de Souza*
- 94-07 **Interfaces Homem-Computador: Uma Primeira Introdução**, *Fábio Nogueira de Lucena e Hans K. E. Liesenberg*
- 94-08 **Reasoning about another agent through empathy**, *Jacques Wainer*
- 94-09 **A Prolog morphological analyser for Portuguese**, *Jacques Wainer, Alexandre Farcic*
- 94-10 **Introdução aos Estadogramas**, *Fábio N. de Lucena, Hans K. E. Liesenberg*
- 94-11 **Matching Covered Graphs and Subdivisions of K_4 and \overline{C}_6** , *Marcelo H. de Carvalho and Cláudio L. Lucchesi*
- 94-12 **Uma Metodologia de Especificação de Times Assíncronos**, *Hélvio Pereira Peixoto, Pedro Sérgio de Souza*

Relatórios Técnicos – 1995

- 95-01 **Paradigmas de algoritmos na solução de problemas de busca multidimensional**, *Pedro J. de Rezende, Renato Fileto*
- 95-02 **Adaptive enumeration of implicit surfaces with affine arithmetic**, *Luiz Henrique de Figueiredo, Jorge Stolfi*
- 95-03 **W3 no Ensino de Graduação?**, *Hans Liesenberg*
- 95-04 **A greedy method for edge-colouring odd maximum degree doubly chordal graphs**, *Celina M. H. de Figueiredo, João Meidanis, Célia Picinin de Mello*
- 95-05 **Protocols for Maintaining Consistency of Replicated Data**, *Ricardo Anido, N. C. Mendonça*
- 95-06 **Guaranteeing Full Fault Coverage for UIO-Based Methods**, *Ricardo Anido and Ana Cavalli*
- 95-07 **Xchart-Based Complex Dialogue Development**, *Fábio Nogueira de Lucena, Hans K.E. Liesenberg*
- 95-08 **A Direct Manipulation User Interface for Querying Geographic Databases**, *Juliano Lopes de Oliveira, Claudia Bauzer Medeiros*
- 95-09 **Bases for the Matching Lattice of Matching Covered Graphs**, *Cláudio L. Lucchesi, Marcelo H. Carvalho*
- 95-10 **A Highly Reconfigurable Neighborhood Image Processor based on Functional Programming**, *Neucimar J. Leite, Marcelo A. de Barros*
- 95-11 **Processador de Vizinhança para Filtragem Morfológica**, *Ilka Marinho Barros, Roberto de Alencar Lotufo, Neucimar Jerônimo Leite*
- 95-12 **Modelos Computacionais para Processamento Digital de Imagens em Arquiteturas Paralelas**, *Neucimar Jerônimo Leite*
- 95-13 **Modelos de Computação Paralela e Projeto de Algoritmos**, *Ronaldo Parente de Menezes e João Carlos Setubal*
- 95-14 **Vertex Splitting and Tension-Free Layout**, *P. Eades, C. F. X. de Mendonça N.*
- 95-15 **NP-Hardness Results for Tension-Free Layout**, *C. F. X. de Mendonça N., P. Eades, C. L. Lucchesi, J. Meidanis*
- 95-16 **Agentes Replicantes e Algoritmos de Eco**, *Marcos J. C. Euzébio*
- 95-17 **Anais da II Oficina Nacional em Problemas Combinatórios: Teoria, Algoritmos e Aplicações**, *Editores: Marcus Vinicius S. Poggi de Aragão, Cid Carvalho de Souza*

- 95-18 **Asynchronous Teams: A Multi-Algorithm Approach for Solving Combinatorial Multiobjective Optimization Problems**, *Rosiane de Freitas Rodrigues, Pedro Sérgio de Souza*
- 95-19 **wxWindows: Uma Introdução**, *Carlos Neves Júnior, Tallys Hoover Yunes, Fábio Nogueira de Lucena, Hans Kurt E. Liesenberg*
- 95-20 **John von Neumann: Suas Contribuições à Computação**, *Tomasz Kowaltowski*
- 95-21 **A Linear Time Algorithm for Binary Phylogeny using PQ-Trees**, *J. Meidanis and E. G. Munuera*

Departamento de Ciência da Computação — IMECC
Caixa Postal 6065
Universidade Estadual de Campinas
13081-970 – Campinas – SP
BRASIL
`reltec@dcc.unicamp.br`