Estudo de Corpus para Classificação de Expressões Anafóricas da Língua Portuguesa

J. C. B. Coelho, Sandra Collovini, Renata Vieira

Programa Interdisciplinar de Pós-Graduação em Computação Aplicada (PIPCA) – Universidade do Vale do Rio dos Sinos (UNISINOS)

Av. Unisinos, 950 – 93.022-000 – São Leopoldo, RS – Brasil

{ccoelho}@icaro.unisinos.br, {sandrac, renata}@exatas.unisinos.br

Abstract. *This work presents a Portuguese corpus study of definite descriptions and its anaphoric and non anaphoric uses. This study serves as the basis to the development of a classification and anaphora resolution system.*

Resumo. *Este trabalho apresenta um estudo de corpus da Língua Portuguesa sobre as descrições definidas e seus usos anafóricos e não anafóricos. Este estudo serve de base ao desenvolvimento de um sistema de classificação e resolução de anáforas.*

1. Introdução

Este trabalho apresenta um estudo de corpus sobre as descrições definidas e seus usos anafóricos e não anafóricos. Para estudar as descrições definidas um conjunto de textos foi anotado automaticamente com informações sintáticas e, depois, manualmente com informações sobre as relações anafóricas.

Este estudo serve de base ao desenvolvimento de um sistema de classificação e resolução de anáforas. Com esse objetivo final, investigamos características que poderiam estar associadas às descrições definidas não anafóricas. Concentramos nossos esforços nesta classe pelo seu número expressivo no corpus. Estudos anteriores confirmam esse dado [Vieira et al, 2003b]. Para avaliar as características levantadas, é apresentada uma classificação automática com Árvore de Decisão.

O trabalho encontra-se assim organizado: na seção 2, a classificação adotada e as concepções ligadas ao processo de referenciação – correferência, anáforas, descrições definidas – são abordadas. Na seção 3, descrevemos o corpus utilizado e suas respectivas anotações. Na seção 4, um estudo das características das descrições definidas não anafóricas é apresentado. Na seção 5, a indução automática de um classificador utilizando essas características, sua avaliação e uma análise de erros são expostas. Reservamos a seção 6 às considerações finais.

2. Referenciação e expressões anafóricas

Considera-se, freqüentemente, a referência (referenciação) como a propriedade da linguagem de representar classe de coisas, pessoas, animais, lugares, fatos etc. Essas instruções lingüísticas usadas para designar são denominadas expressões referenciais.

A organização dessas expressões referenciais é fundamental à continuidade e à estabilidade do texto. Esses dois conceitos também são conhecidos, respectivamente, como progressão temática e coesão textual [Conte, 1996; Francis, 1994; Koch, 2002; Schwarz, 2000]. Basicamente, nesse processo de organização utilizamos termos/expressões que retomam outros termos do próprio texto, constituindo, assim, cadeias referenciais. Esse fenômeno de retomada de referentes, é denominado correferência. No exemplo:

> (1) *O Eurocenter oferece cursos de Japonês na bela cidade de Kanazawa. Os cursos têm quatro semanas de duração. As aulas do nível avançado incluem refeições típicas e passeios a pontos turísticos*.

observamos que na frase inicial são introduzidas três referentes (*O Eurocenter, cursos de Japonês* e *a bela cidade de Kanazawa*). Na frase seguinte (na continuidade do texto), a expressão "*Os cursos*" retoma "*cursos de Japonês*". Nessa perspectiva, "*cursos de Japonês*" é antecedente de "*Os cursos*", ou seja, duas expressões referenciais que fazem menção à mesma entidade, portanto duas expressões correferenciais.

No contexto da referenciação, além do conceito de correferência, encontramos a noção de anáfora. Tradicionalmente, a anáfora se define como toda retomada de um elemento anterior em um texto, mantendo-se a identidade referencial. Atualmente essa noção vem sendo ampliada. Reconhece-se que a anáfora não é necessariamente correferencial e que o referente de uma expressão anafórica não é sempre explicitamente denotado por um mesmo referente anterior [McEnery, 1998]. No exemplo (1), observamos que a expressão "*As aulas do nível avançado*" não é correferente a nenhum termo anterior, mas apresenta parte do seu significado apoiado da expressão "*cursos de Japonês*". Assim, a anáfora é um fenômeno de semântica textual de natureza inferencial, podendo ser ou não um fenômeno de correferência. Nesse sentido, uma expressão anafórica pode retomar uma referência anterior, mas também pode ativar um novo referente cuja interpretação é dependente de outras expressões referenciais anteriormente presentes do texto (seção 4).

2.1. Descrições definidas

A referenciação, e conseqüentemente a correferência e a anáfora, pode realizar-se por intermédio de formas gramaticais que exercem a função de pronome ou por intermédio de sintagmas[1] nominais.

Devido à variedade de expressões referenciais, concentramos nossos esforços em sintagmas nominais definidos, denominados descrições definidas. Essas são compostas por um nome núcleo e iniciadas por um artigo definido (*o, a, os, as*) acompanhado, por vezes, de elementos complementares. As descrições definidas, diferentemente das formas pronominais, podem tanto retomar referencias do texto (anafórica) quanto introduzirem novos referentes (não anafórica). Em (1), um exemplo de descrição definida anafórica é "*Os cursos*" e um exemplo de descrição definida não anafórica é "*O Eurocenter*".

[1] Sintagma é a unidade da análise sintática composta de um núcleo e de outros termos que a ele se unem, formando uma locução que entrará na formação da oração. O nome do sintagma depende da classe da palavra que forma seu núcleo, havendo assim sintagma nominal (núcleo nome), sintagma adjetivo (núcleo adjetivo), sintagma preposicional (núcleo preposição) etc.

2.2. Classificação das descrições definidas

Em Vieira (1998), encontra-se uma divisão das descrições definidas em 4 classes, dependendo da forma que estão relacionadas com os seus antecedentes. A seguir, são apresentadas as classes dessa taxionomia com exemplos do corpus:

- **Correferentes**
  - **Anafóricas diretas**: são antecedidas por uma expressão que se refere à mesma entidade no discurso com mesmo nome núcleo. Por exemplo: "*As listas apontam quase todas as divisões e departamentos da Polícia Civil. Alguns delegados federais também são citados nas listas*".
  - **Anafóricas indiretas**: são antecedidas por uma expressão que se refere à mesma entidade no discurso com nome núcleo diferente. Por exemplo: "*A Folha de São Paulo apresentou as listas apreendidas na operação contra o crime organizado. O jornal tentou ouvir o delegado encarregado*".
- **Não correferentes**
  - **Anafóricas associativas**: introduz um novo referente que possui parte do seu significado ancorado em uma expressão anterior. Por exemplo: "*A Folha de São Paulo apresentou as listas apreendidas na operação contra o crime organizado. O jornal tentou ouvir o delegado encarregado*".
  - **Não anafóricas**: são aquelas que introduzem um novo referente no texto e não possuem uma âncora para se apoiar semanticamente. Por exemplo: "*O quilômetro 430 da rodovia Assis Chateau Briand ontem foi cenário da campanha de segurança no trânsito*".

3. Anotação do corpus

3.1 Corpus

O estudo apresentado aqui foi realizado em um corpus constituído por 24 textos jornalísticos da Folha de São Paulo (um extrato do corpus NILC[2]), escritos em português do Brasil.

Tabela 1. Informações sobre o corpus

N.º Textos	Tamanho Textos	N.º Palavras	Nº Sintagmas Nominais	Nº Descrições Definidas
24	de 1 a 6 Kb	11042	2319	1411

3.2. Anotação automática

O corpus foi anotado automaticamente com informações sintáticas e semânticas. Para isso, foi utilizado um analisador sintático do português, o PALAVRAS [Bick, 2000]. Uma segunda ferramenta foi empregada, Palavras Xtractor [Gasperin et al, 2003]. Esta ferramenta converte a saída do analisador sintático PALAVRAS em três arquivos XML[3] (*EXtensible Markup Language*): **i)** *words* (com uma lista de palavras do texto e seus respectivos identificadores), **ii)** *pos* (com as informações sintáticas e semânticas das palavras do texto) e **iii)** *chunks* (com a estrutura do texto).

[2] Núcleo Interinstitucional de Lingüística Computacional. Disponível em http://www.nilc.icmp.usp.br/nilc

[3] Disponível em: http://www.w3.org/XML.

3.3. Anotação manual

As descrições definidas do corpus foram anotadas manualmente com informações sobre as relações anafóricas. Para isso, utilizamos uma ferramenta de anotação de correferência no discurso, MMAX (*Multi-Modal Annotation in XML*) [Müller and Strube, 2000]. A anotação feita com esta ferramenta segue a metodologia *stand-off,* em que a marcação é armazenada separadamente do corpus. O resultado do processo de anotação no MMAX é um arquivo XML de marcações (*markables*), sendo que cada item marcado contém atributos cujos valores são definidos pelo projetista do estudo de corpus, em nosso caso, a classificação apresentada na seção 2.2.

A anotação manual do corpus ocorreu em 4 passos: **i)** seleção das descrições definidas; **ii)** classificação das descrições definidas em *correferentes* ou *não correferentes*, sendo que para as correferentes foram apontados os antecedentes; **iii)** classificação das descrições definidas correferentes em *anafóricas diretas* ou *indiretas*; **iv)** classificação das descrições definidas não correferentes em *não anafóricas* ou *anafóricas associativas*, sendo que para as anafóricas associativas foram apontadas as expressões na qual parte do sentido estava ancorada [Vieira et al, 2003a]. O corpus foi anotado por 2 anotadores. As marcações que não entraram em consenso foram lançadas na classe *outra* (Tabela 2).

Tabela 2. Resultado da anotação manual do corpus

Classes	Totais	Classes	Totais
correferente	515	anafórica direta	364
		anafórica indireta	151
não correferente	816	não anafórica	696
		anafórica associativa	120
outra			80
		Total Geral	**1411**

As marcações que não entraram em consenso (classe outra), ocorreram, predominantemente, entre a classe não anafórica e anafórica associativa – aproximadamente 62 casos. Ao contrário das classes correferentes anafórica direta e indireta, que possuem uma marca de distinção estrutural (o nome núcleo é igual no primeiro caso e diferente no segundo caso), as classes não correferentes anafórica associativa e não anafórica apresentam um tênue traço de distinção semântico (ambas introduzem novos referentes no discurso, porém as anafóricas associativas têm parte do sentido ancorado em uma expressão anterior no texto).

4. Estudo das características das descrições definidas não anafóricas

De posse do corpus anotado, foi verificado que o número de descrições definidas não anafóricas era expressivo (Tabela 2). Estudos anteriores confirmam esse dado [Vieira et al, 2003b]. Realizamos, então, um estudo das descrições definidas com o objetivo de investigar características que poderiam estar associadas à classe não anafórica. Na literatura, encontramos várias propostas para classificação automática de descrições definidas [Muller et al., 2002; Poesio et al., 2005; Uryupina, 2003]. A partir desses trabalhos, realizados na grande maioria para o inglês, e de um estudo inicial para o português [Collovini, 2004], organizamos 16 características para classificação de descrições definidas não anafóricas para o português. Os conceitos lingüísticos empregados nesta seção (e.g. aposto, superlativo, cláusula relativa etc.) estão em acordo com (Neves, 2000; Vilela and Koch, 2001).

Os exemplos de descrições definidas do corpus foram submetidos a uma minuciosa análise quanto a sua estrutura sintática. Predominantemente, observamos que a primeira manifestação de um referente na forma de descrição definida (não anafóricas) ocorre de modo mais completo do que as demais. Por isso, os elementos complementares – elementos que se integram ao nome núcleo para completá-lo ou aperfeiçoá-lo – foram selecionados como características da classe não anafórica: **SP** (descrições definidas formadas por sintagma preposicional, e.g., "*A tarde de ontem*"); **REL** (descrições definidas formadas por cláusula relativa, ou seja, unidades encaixadas por meio de um pronome relativo, e.g., "*O texto que deve ser assinado pelos jornalistas*"); **SA** (descrição definida formada por um sintagma adjetival posposto ao nome núcleo, e.g., "*As conversas mais antigas*"); **PRE_ADJ** (descrição definida que apresenta um adjetivo anteposto ao núcleo, e.g., "*O primeiro grau*"); **PRE_NUM** (descrição definida composta por um numeral anteposto ao núcleo, e.g., "*Os 65 anos*"); **NUM** (descrição definida que apresenta após o núcleo um numeral, e.g., "*Os anos 60*"); **DET** (descrição definida que possui, além do artigo definido, outro(s) determinante(s) como pronomes indefinidos, possessivos e demonstrativos, e.g., "*Os nossos arqueólogos*"); **SUP** (descrição definida composta por superlativo, ou seja, termo que expressa uma qualidade em um grau elevado, e.g., "*Os melhores alunos*") e **SUP_A** (descrição definida que descreve o grau máximo de qualidade (adjetivo), representando o maior índice de uma escala, e.g., "*O Christofle líquido é o melhor*").

Também, ligado ao caráter completo da primeira manifestação de um referente, verificamos que as estruturas com nomes próprios compostos coincidem, predominantemente, com a classe não anafórica. Geralmente, na continuidade do texto, as expressões que retomam descrições definidas com nomes próprios compostos o fazem com apenas um dos nomes próprios ou com termos sinônimos. Por exemplo, a primeira manifestação do referente "*o Stbstudent Travel Bureau*" é retomado pelos termos "*o Stbstudent*" e "*a agência*". As características resultantes dessa restrição foram: **NP_COM** (descrição definida com nome núcleo próprio composto, e.g., "*O Othon Palace Hotel*"); **APO** (descrições definidas com construção de aposto[4] com marca explícita, como vírgulas ou travessões, e.g., "*O prefeito de Gravataí, Daniel Luiz Bordignom*") e **APO_NP** (descrição definida composta por nome próprio constituindo um aposto sem marca explícita. Por exemplo: "*O delegado Elson Campelo*"). Ao examinar, todas as características até então levantadas, constatamos que o número de palavras que formam as descrições definidas poderia ser fator determinante para os casos não anafóricos. Assim foi elaborada a característica **TAM** (descrição definida formada por cinco ou mais termos, e.g., "*As mais recentes criações estéticas brasileiras*"). Essas características baseadas na estrutura da descrição definidas foram reunidas para compor nosso primeiro grupo de características (G1).

A literatura destaca o caráter não anafórico das relações que ocorrem, em descrições definidas, com verbos de ligação (ser, estar, permanecer etc) denominadas *construções copulares* (Vieira, 1998; Poesio et al., 2005). Esse dado vai ao encontro da concepção apresentada no parágrafo anterior – composições mais completas sinalizam a primeira manifestação de um referente no texto. Por isso, foi elaborada a característica: **COP** (descrições definidas em uma construção copular, e.g., "*O coreano seria a língua dos anjos*"). Entendendo que descrições definidas não anafóricas não apresentam

[4] Um ou mais termos que se referem a um substantivo ou pronome explicando-o.

antecedentes e que, geralmente, as expressões da primeira sentença não possuem antecedentes, elaboramos uma característica que analisa a posição da sentença: **PRI_SENT** (descrições definidas no título ou na primeira sentença do texto). Essas duas características formam o segundo grupo (G2), que envolvem unicamente a análise da sentença.

Na coesão textual, são utilizadas expressões que retomam outras do próprio texto. Isso, por vezes, ocorre por meio de repetições dos nomes núcleos. Esses casos são tão freqüentes (364 casos no corpus, 71% dos casos correferentes) que constituem uma classe específica, as anafóricas diretas. Seguindo esse dado foi desenvolvida, então, a característica **SEM_ANT** (o núcleo da descrição definida é uma palavra que não ocorre anteriormente no texto). Esta compõe o terceiro grupo (G3).

As características levantadas foram organizadas em três grupos a fim de podermos examinar melhor o desempenho de cada aspecto.

5. Classificação automática

Muitas aplicações de Processamento da Linguagem Natural requerem o tratamento de correferência. No contexto desse trabalho, os grupos de características (G1-2-3) apresentados na seção 4, foram utilizados como atributos para a classificação das descrições definidas com Árvores de Decisão[5] (Quinlan, 1993). Para isso foi utilizada a ferramenta Weka (*Waikato Environment for Knowledge Analysis*) [Witten and Frank, 2000] que implementa o algoritmo *j48*[6]. Na classificação automática foi utilizado o método de validação cruzada (*cross-validation*) em *10 folds*[7].

Foram realizadas três classificações automáticas distintas. Foi realizada uma classificação considerando 4 classes (não anafórica, anafórica associativa, anafórica direta e anafórica indireta – veja seção 2.2), porém não houve distinção das classes anafórica indireta e associativa. A segunda classificação foi feita entre as classes *"não_correferente"* (i.e. anafóricas associativas e não anafóricas) e *"outra"* (i.e. as demais classificações). A terceira foi executada entre as classes *"não_anaforica"* e *"outra*" (i.e. as demais classificações). As Árvores de Decisão geradas são ilustradas nas figuras: Figura 1, Figura 2, Figura 3, Figura 4, Figura 5 e Figura 6.

Nas Árvores de Decisão geradas com atributos G1 (Figura1 e Figura2), são comuns os atributos: TAM, SA, NP_COM, PRE_ADJ. A Árvore de Decisão (Figura 2), além desses atributos, apresenta também os atributos APO_NP e SP. Observamos que os atributos G1, que consideram unicamente a estrutura do sintagma, apresentam resultados consideráveis de precisão em relação aos *baselines* que consideram todas as expressões *não_anaforica* ou *não_correferente*. Para classe *"não_anaforica"* obtivemos 65% de precisão em contraponto a 50% do *baseline*. Na classe *"não_correferente"*, encontramos 71% em relação a 60% do *baseline* (Tabela 3).

---

[5] Árvores de Decisão é uma técnica de classificação com base em atributos e podem ser geradas a partir de uma base de dados previamente classificada para serem posteriormente utilizadas na classificação de novos exemplos.

[6] O algoritmo *j48* é uma re-implementação em Java do algoritmo de Árvores de Decisão *C4.5*.

[7] Validação Cruzada em *10 folds*: a base de dados é dividida em 10 partes, sendo 9 partes para o treinamento e a parte remanescente para teste do algoritmo de classificação. Esse processo é repetido 10 vezes, cada vez considerando uma parte diferente para teste.

```
TAM = TRUE: não_anaforica (271.0/87)
TAM = FALSE
| SA = TRUE: não_anaforica (94.0/32)
| SA = FALSE
| | NP_COM = TRUE: não_anaforica (60/23)
| | NP_COM = FALSE
| | | PRE_ADJ = TRUE: não_anaforica (24/6)
| | | PRE_ADJ = FALSE: outra (656/249)
```

Figura 1. AD - classe *"não_anaforica"* e atributos G1

```
TAM = TRUE: nao_correferente (271/70)
TAM = FALSE
| NUM = TRUE: nao_correferente (6/1)
| NUM = FALSE
| | APO_NP = TRUE: nao_correferente (33/8)
| | APO_NP = FALSE
| | | SA = TRUE: nao_correferente (92/29)
| | | SA = FALSE
| | | | NP_COM = TRUE: nao_correferente (55/19)
| | | | NP_COM = FALSE
| | | | | PRE_ADJ = TRUE: nao_correferente (21/4)
| | | | | PRE_ADJ = FALSE
| | | | | | SP = TRUE: nao_correferente (46/18)
| | | | | | SP = FALSE: outra (581/269)
```

Figura 2. AD - classe *"não_correferente"* e atributos G1

```
PRI_SENT = TRUE: não_anaforica (35)
PRI_SENT = FALSE
| TAM = TRUE
| | APO_NP = TRUE: outra (3/1)
| | APO_NP = FALSE: não_anaforica (258/85)
| TAM = FALSE
| | SA = TRUE: não_anaforica (91/32)
| | SA = FALSE
| | | PRE_ADJ = TRUE: não_anaforica (24/6)
| | | PRE_ADJ = FALSE
| | | | NP_COM = TRUE: não_anaforica (56/23)
| | | | NP_COM = FALSE
| | | | | APO_NP = TRUE: não_anaforica (24/11)
| | | | | APO_NP = FALSE: outra (614/218)
```

Figura 3. Árvore de Decisão da classe *"não_anaforica"* com atributos G1-2

```
PRI_SENT = TRUE: nao_correferente (39)
PRI_SENT = FALSE
| TAM = TRUE: nao_correferente (261/70)
| TAM = FALSE
| | APO_NP = TRUE: nao_correferente (32/8)
| | APO_NP = FALSE
| | | NUM = TRUE: nao_correferente (5/1)
| | | NUM = FALSE
| | | | SA = TRUE: nao_correferente (89/29)
| | | | SA = FALSE
| | | | | NP_COM = TRUE: nao_correferente (52/19)
| | | | | NP_COM = FALSE
| | | | | | PRE_ADJ = TRUE: nao_correferente (21/4)
| | | | | | PRE_ADJ = FALSE
| | | | | | | SP = TRUE: nao_correferente (46/18)
| | | | | | | SP = FALSE: outra (560/248)
```

Figura 4. Árvore de Decisão da classe *"não_correferente"* com atributos G1-2

```
PRI_SENT = TRUE: não_anaforica (35)
PRI_SENT = FALSE
| SEM_ANT = TRUE: não_anaforica (705/260)
| SEM_ANT = FALSE
| | SUP = TRUE: não_anaforica (6/1)
| | SUP = FALSE
| | | NUM = TRUE: não_anaforica (4/1)
| | | NUM = FALSE
| | | | SA = TRUE
| | | | | COP = TRUE
| | | | | | SP = TRUE: outra (3/1)
| | | | | | SP = FALSE: não_anaforica (4/1)
| | | | | COP = FALSE
| | | | | | TAM = TRUE: não_anaforica (4/1)
| | | | | | TAM = FALSE: outra (24/9)
| | | | SA = FALSE: outra (320/46)
```

Figura 5. Árvore de Decisão da classe *"não_anaforica"* com atributos G1-2-3

```
SEM_ANT = TRUE: nao_correferente (737/174)
SEM_ANT = FALSE
| NUM = TRUE: nao_correferente (5)
| NUM = FALSE
| | PRI_SENT = TRUE: nao_correferente (3)
| | PRI_SENT = FALSE
| | | SUP = TRUE: nao_correferente (6/1)
| | | SUP = FALSE: outra (354/68)
```

Figura 6. Árvore de Decisão da classe *"não_correferente"* com atributos G1-2-3

Nas Árvores de Decisão das Figuras 4 e 5, destacam-se os atributos: PRI_SENT, TAM, APO_NP, SA, PRE_ADJ, NP_COM. Na Figura 4 estão presentes também os atributos NUM e SP. Do grupo de atributos G2, o atributo PRI_SENT foi incluído no topo das árvores, já o atributo COP não foi considerado. A combinação de atributos G1-2 apresentou o melhor resultado em precisão (66%) para a classe *"não_anaforica"*. Já a classe *"não correferente"* com atributos G1-2 não apresentou diferença significativa em relação aos atributos G1. Os atributos presentes nas Árvores de Decisão Figura 5 e Figura 6 são: PRI_SENT, SEM_ANT, SUP, NUM. Os atributos SA, COP, SP, TAM aparecem na árvore da Figura 5 (classe *"não_anaforica"*). Os atributos do grupo G2 (PRI-SENT e COP) foram considerados nessa árvore; e o novo atributo do grupo G3 passou a substituir, em ambas as árvores, vários atributos dos demais grupos. Como resultados, a combinação de atributos G1-2-3 apresentou para a classe *"não_anaforica"* 65% de precisão e para a classe *"não_correferente"* 76% de precisão.

De forma geral, destacou-se a combinação de atributos G1-2-3. De acordo com a Tabela 3, observamos um ganho em abrangência com os atributos G1-2-3 em relação aos demais (55% com G1, 57% com G1-2 e 88% com G1-2-3). Para a classe *"não_correferente"* a combinação de atributos G1-2-3 apresentou um pequeno aumento em precisão (76%) em relação aos demais grupos (71%). Também, foi possível observar um ganho significativo em abrangência com os atributos G1-2-3 (58% com G1, 61% com G1-2 e 89% com G1-2-3).

Tabela 3. Classificação das descrições definidas

	Classes	G1	G1-2	G1-2-3
Abrangência	não_anaforica	55%	57%	88%
	não_correferente	58%	61%	89%
Precisão	não_anaforica	65%	66%	65%
	não_correferente	71%	71%	76%

Na análise dos erros, observamos que algumas descrições definidas anafóricas associativas foram classificadas como não anafóricas. Na maioria dos casos, as descrições definidas que têm elementos complementares na sua composição são não anafóricas, porém há exceções. Por exemplo, são não anafóricas as descrições definidas "*a carreira do magistério*" e "*A revolução tecnológica*"; essas apresentam respectivamente os complementos sintagma preposicional e sintagma adjetival, e, conseqüentemente, foram classificadas corretamente pelas características **SP** e **SA**. Contudo, estas duas características classificaram incorretamente duas retomadas incompletas de sentido sob a forma de descrições definidas anafóricas associativas: "*as propostas em discussão*" (**SP**), "*o sistema básico*" (**SA**).

De forma parecida, na grande maioria dos casos, as descrições definidas anafóricas diretas, indiretas e associativas não têm elementos complementares, contudo, foram identificadas exceções no corpus. Por exemplo, as descrições definidas "*os cursos*" e "*o trecho*" não apresentam elementos complementares e, por isso, foram classificadas corretamente como anafóricas – seus antecedentes são respectivamente "*os cursos de japonês*" e "*O quilômetro 430 da rodovia Assis Chateau Briand*". No entanto, as descrições definidas "*a justiça*" e "*a ética*" não possuem elementos complementares e são não anafóricas. Isso ocorre porque seu sentido se constrói num conhecimento compartilhado pelos leitores.

6. Considerações Finais

Neste trabalho foi apresentado um estudo de corpus para verificar características das descrições definidas que possam sinalizar quando estas são anafóricas ou não anafóricas. O estudo das características baseou-se na literatura e no estudo de corpus realizado. Foi construída uma base de dados para a classificação automática com Árvores de Decisão.

Em relação aos erros, percebemos, na classificação automática, uma dificuldade também presente na anotação manual: a classe anafórica associativa e não anafórica apresentam maiores dificuldades de dissociação do que as demais classes, uma vez que ambas introduzem novos referentes no discurso. Contudo as descrições definidas anafóricas associativas ancoram seu significado em uma expressão anterior. A identificação das relações semânticas que permitem esse ancoramento é uma tarefa difícil do ponto de vista computacional.

No entanto, como o analisador sintático PALAVRAS começou recentemente a incorporar marcações semânticas no resultado da análise, uma solução seria incorporar essa informação. Pode-se comparar traços semânticos dos nomes núcleos das descrições definidas com de expressões antecedentes. Assim, traços semânticos equivalentes poderiam indicar a anáfora associativa. Por exemplo, se aplicarmos o princípio semântico sugerido, observaremos no trecho "*Cursos ensinam o japonês e modo de vida. Após um mês, aluno aplicado entenderá e falará modestamente a língua*" que a descrição definida "*a língua*" apresenta uma relação com a descrição definida "*o japonês*", uma vez que ambas possuem o traço semântico (marcado automaticamente) "*ling*", referente à língua humana. Logo, a descrição definida "*a língua*" seria classificada automaticamente como anafórica associativa. Como trabalhos futuros estamos desenvolvendo essa solução.

Agradecimentos

Este artigo foi parcialmente realizado com apoio da Capes, CNPq, FAPERGS e UNISINOS.

Referências

Bick, E. (2000) "The Parsing System PALAVRAS: Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework". PhD thesis, Arhus University, Arhus.

Collovini, S.; Goulart, R.; Vieira, R. (2004) "Identificação de Expressões Anafóricas e Não Anafóricas com Base na Estrutura do Sintagma". In: 2.º Workshop em Tecnologia da Informação e da Linguagem Humana (TIL 2004) – Salvador, BA.

Conte, Elisabeth (1996) "Anaphoric encapsulation". Belgian Journal of Linguistics: Coherence and anaphora, v.10, p.1-10.

Francis, Gill (1994) "Labelling discurse: an aspect of nominal-group lexical cohesion". In Coulthard, Malcon (ed.). Advances in written text analysis. Londres: Routledge.

Gasperin, C.; Vieira, R.; Goulart, R.; Quaresma, P. (2003) "Extrating XML Syntactic Chunks from Portuguese Corpora". In: Traitement Automatique Dês Langues Minoritaires- TALN, Btaz-sur-mer, France.

Koch, Ingedore G. V. (2002) “Desvendando os segredos do texto”. São Paulo: Cortez.

McEnery,T.; Botley, S (1998) (Eds) Discourse Anaphora and Anaphor Resolution. Amsterdam, John Benjamins.

Müller, C. and Strube, M. (2000) “MMAX: A tool for the annotation of multi-modal corpora”. In: Proceedings of the IJCAI 2001, Seattle, p. 45–50.

Müller, C.; Stefan, R.; Strube, M. (2002) “Applying Co-training to reference resolution”. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL- 2002), Philadelphia, Penn., p. 352-359.

Neves, Maria Helena de Moura (2000) “Gramática de usos do português”. São Paulo: UNESP - Universidade Estadual Paulista.

Poesio, M.; Alexandrov-Ksbadjov, M.; Vieira, R.; Goulart, R.; Uryupina, O. (2005) “Does Discourse-new Detection Help Definite Description Resolution?”. In: Sixth International Workshop on Computational Semantics (IWCS 6), Tiburg.

Quinlan, J.R. (1993) “C4.5: Programs for Machine Learning”. Morgan Kaufmann, San Mateo, CA.

Santos, D. (2000) “O projecto Processamento Computacional do Português: Balanço e Perspectivas”. In: V Encontro para o Processamento da Língua Portuguesa Escrita e Falada (PROPOR). Atibaia, São Paulo, Brasil, p. 105-113.

Schwarz, Monika (2000) Indirekte Anaphen in Texten. Tübingen: Niemeyer.

Soon, W. M.; Ng, H.wee T.; Lim, D. C. Y. (2001) “A machine learning approach to coreference resolution of noun phrases”. In: Computational Linguistics, p. 521–544.

Uryupina, O. (2003) “High-precision Identification of Discourse New and Unique Noun Phrases”. In: Proceedings of the ACL Student Workshop, Sapporo.

Vieira, R. (1998) “Definite description processing in unrestricted text”. PhD thesis, University of Edinburgh, Edinburgh.

Vieira, R.; Gasperin, C.; Goulart, R. (2003a) “From manual to automatic annotation of coreference”. In Proceedings of the International Symposium on Reference Resolution and Its Applications to Question Answering and Summarization, Venice.

Vieira, R.; Gasperin, C.; Goulart, R.; Salmon-Alt, S. (2003b) “From concrete to virtual annotation mark-up language: the case of COMMOn-REFs”. In Proceedings of the (ACL 2003) Workshop on Linguistic Annotation: Getting the Model Right, Sapporo.

Vilela, Mário; Koch, Ingedore Grunfeld Villaça (2001) “Gramática da língua portuguesa: gramática da palavra, gramática da frase, gramática do texto/discurso.” Coimbra: Almedina.

Witten, Ian H.; Frank, Eibe (2000) “Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations”. Morgan Kaufmann Publishers.